Agreement between Apple Watch and Actical step counts in a community setting: The Framingham Heart Study

Nicole L Spartano, Yuankai Zhang, Chunyu Liu, Ariel Chernofsky, Honghuang Lin, Ludovic Trinquart, Belinda Borrelli, Chathurangi Heshani Pathiravasan, Vik Kheterpal, Christopher Nowak, Ramachandran S Vasan, Emelia J Benjamin, David D McManus, Joanne M Murabito

Submitted to: JMIR Biomedical Engineering
on: November 16, 2023

Table of Contents

Agreement between Apple Watch and Actical step counts in a community setting: The Framingham Heart Study

Nicole L Spartano[1, 2] PhD; Yuankai Zhang[3] MS; Chunyu Liu[3] PhD; Ariel Chernofsky[3] PhD; Honghuang Lin[4] PhD; Ludovic Trinquart[5, 6] PhD, MPH; Belinda Borrelli[7] PhD; Chathurangi Heshani Pathiravasan[3] PhD; Vik Kheterpal[8] MD; Christopher Nowak[8] PhD; Ramachandran S Vasan[9, 10, 11, 12, 13] MD; Emelia J Benjamin[10, 11, 2] MD, ScM; David D McManus[14, 15] MD, ScD; Joanne M Murabito[2, 16] MD, ScD

[1]Section of Endocrinology, Diabetes, Nutrition, and Weight Management, Boston University Chobanian and Avedisian School of Medicine Boston US
[2]Boston University's and National Heart, Lung, and Blood Institute's Framingham Heart Study Framingham US
[3]Boston University School of Public Health Boston US
[4]University of Massachusetts Chan Medical School Worcester US
[5]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center Boston US
[6]Tufts Clinical and Translational Science Institute, Tufts University Boston US
[7]Boston University Henry M. Goldman School of Dental Medicine Boston US
[8]Care Evolution Ann Arbor US
[9]Section of Preventive Medicine and Epidemiology, Department of Medicine Boston University Chobanian and Avedisian School of Medicine Boston US
[10]Section of Cardiology, Department of Medicine Boston University Chobanian and Avedisian School of Medicine Boston US
[11]Department of Epidemiology Boston University Chobanian and Avedisian School of Medicine and Boston University School of Public Health Boston US
[12]University of Texas School of Public Health and University of Texas Health Sciences Center San Antonio US
[13]1. Boston University's and National Heart, Lung, and Blood Institute's Framingham Heart Study Framingham US
[14]Department of Medicine, University of Massachusetts Chan Medical School Worcester US
[15]Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School Worcester US
[16]Section of General Internal Medicine, Department of Medicine, BUCASM and Boston Medical Center, Boston University Chobanian and Avedisian School of Medicine and Boston Medical Center Boston US

Corresponding Author:
Nicole L Spartano PhD
Section of Endocrinology, Diabetes, Nutrition, and Weight Management, Boston University Chobanian and Avedisian School of Medicine
72 E Concord St, Suite 301 Collamore
Boston
US

Abstract

Background: Step counting is comparable among many research-grade and consumer-grade accelerometers in laboratory settings, but few studies have compared step count measurement among devices outside of the laboratory, in a community setting.

Objective: The purpose of this study was to compare agreement between Actical and Apple Watch step-counting in a community setting.

Methods: Among Third Generation Framingham Heart Study participants (n=3486), we examined agreement of step-counting between those who wore a consumer-grade (Apple Watch Series 0) and research-grade accelerometer (Actical) on the same day(s). Secondarily, we examined agreement during each hour when both devices were worn to account for differences in wear time between devices.

Results: We studied 523 participants (n=3223 person-days, mean age 51.7 years, 57% women). Between devices, we observed modest correlation (intraclass correlation [ICC]=0.56, 95% confidence interval [CI]=0.54, 0.59), poor continuous agreement (29.7% of days having steps counts with ?15% difference), a mean difference of 499 steps/day higher count by Actical, and wide limits of agreement, roughly +/-9000 steps/day. However, devices showed stronger agreement in identifying who meets various

step/day threshold (e.g. at 8000 steps/day, kappa coefficient=0.49), for which devices were concordant for 74.8% of participants. In secondary analyses, of hours during which both devices were worn (456 participants, 18760 person-hours), the correlation was much stronger (ICC=0.86, 95% CI=0.85, 0.86), but continuous agreement remained poor (27.3% of hours having step counts with ?15% difference) between devices and was slightly worse for those with mobility limitations or obesity.

Conclusions: Our investigation suggests poor overall agreement between steps counted by the Actical and Apple Watch devices, with stronger agreement in discriminating who meets certain step thresholds. The impact of these challenges may be minimized if accelerometers are used by individuals to determine whether they are meeting physical activity guidelines or tracking step counts. It is also possible that some of the limitations of these older accelerometers may be improved in newer devices.

(JMIR Preprints 16/11/2023:54631)
DOI: https://doi.org/10.2196/preprints.54631

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Agreement between Apple Watch and Actical step counts in a community setting: The Framingham Heart Study

Nicole L. Spartano PhD,[1,2] Yuankai Zhang MA,[3] Chunyu Liu PhD,[3] Ariel Chernofsky PhD,[3] Honghuang Lin PhD,[4] Ludovic Trinquart PhD MPH,[5,6] Belinda Borrelli PhD,[7] Chathurangi H. Pathiravasan PhD,[3] Vik Kheterpal, MD,[8] Christopher Nowak PhD,[8] Ramachandran S Vasan MD,[1,9,10,11,12] Emelia J. Benjamin MD ScM,[1,10,11] David D. McManus MD ScM,[4,13] Joanne M. Murabito MD ScM[1,14]

1. Boston University's and National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, USA
2. Section of Endocrinology, Diabetes, Nutrition, and Weight Management, Boston University Chobanian and Avedisian School of Medicine (BUCASM), Boston, MA, USA
3. Department of Biostatistics, Boston University School of Public Health (BUSPH), Boston, MA, USA
4. Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, USA
5. Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA
6. Tufts Clinical and Translational Science Institute, Tufts University, Boston, MA, USA
7. Center for Behavioral Science Research, Boston University, Henry M. Goldman School of Dental Medicine, Boston, MA, USA
8. Care Evolution, Ann Arbor, MI, USA
9. Section of Preventive Medicine and Epidemiology, Department of Medicine, BUCASM, Boston, MA, USA
10. Section of Cardiology, Department of Medicine, BUCASM, Boston, MA, USA
11. Department of Epidemiology, BUCASM and BUSPH, Boston, MA, USA
12. UT School of Public Health in San Antonio, TX, and UT Health Sciences Center in San Antonio, TX, USA
13. Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, USA
14. Section of General Internal Medicine, Department of Medicine, BUCASM and Boston Medical Center, Boston, MA, USA

Corresponding Author:
Nicole L. Spartano, PhD
Boston University Chobanian and Avedisian School of Medicine
Section of Endocrinology, Diabetes, Nutrition & Weight Management
72 E. Concord St, Suite C301
Boston, MA 02118
Phone: 315-415-2040
Email: Spartano@bu.edu

Abstract (275 words)

Background: Step counting is comparable among many research-grade and consumer-grade accelerometers in laboratory settings. The purpose of this study was to compare agreement between Actical and Apple Watch step-counting in a community setting.

Methods: Among Third Generation Framingham Heart Study participants (n=3486), we examined agreement of step-counting between those who wore a consumer-grade (Apple Watch Series 0) and research-grade accelerometer (Actical) on the same day(s). Secondarily, we examined agreement during each hour when both devices were worn to account for differences in wear time between devices.

Results: We studied 523 participants (n=3223 person-days, mean age 51.7 years, 57% women). Between devices, we observed modest correlation (intraclass correlation [ICC]=0.56, 95% confidence interval [CI]=0.54, 0.59), poor continuous agreement (29.7% of days having steps counts with ≤15% difference), a mean difference of 499 steps/day higher count by Actical, and wide limits of agreement, roughly +/-9000 steps/day. However, devices showed stronger agreement in identifying who meets various step/day threshold (e.g. at 8000 steps/day, kappa coefficient=0.49), for which devices were concordant for 74.8% of participants. In secondary analyses, of hours during which both devices were worn (456 participants, 18760 person-hours), the correlation was much stronger (ICC=0.86, 95% CI=0.85, 0.86), but continuous agreement remained poor (27.3% of hours having step counts with ≤15% difference) between devices and was slightly worse for those with mobility limitations or obesity.

Conclusion: Our investigation suggests poor overall agreement between steps counted by the Actical and Apple Watch devices, with stronger agreement in discriminating who meets certain step thresholds. The impact of these challenges may be minimized if accelerometers are used by individuals to determine whether they are meeting physical activity guidelines or tracking step counts. It is also possible that some of the limitations of these older accelerometers may be improved in newer devices.

Keywords: accelerometer, mobile health, wearable device, fitness tracker, physical activity

Introduction

Physical inactivity is an important risk factor for many chronic diseases including obesity, diabetes mellitus, hypertension, cardiovascular disease, and dementia.[1] The 2018 Physical Activity Guidelines for Americans recommend 150 minutes of moderate to vigorous physical activity (MVPA) or more per week.[1] Despite many known benefits of physical activity, many Americans do not meet the Physical Activity Guidelines, and national physical activity levels determined using self-report questionnaires may even overestimate true activity levels. Guideline achievement has been estimated to be as low as 15% of Americans using accelerometry in a nationally representative sample, or as high as 66% using self-reported data in the same individuals.[2, 3] Furthermore, experts have expressed concern over whether these guidelines are appropriate and attainable, especially in older adults or those with mobility limitations.[1, 4, 5]

Walking is a central component of physical activity and public health promotion efforts.[6] Public health messages focused on daily step counts may be a more appropriate target for achieving recommended amounts of physical activity in adults,[6] which might have even more significance in older populations and those who have low MVPA levels. We are in a new paradigm in healthcare, in which 69% of US adults report tracking at least one health metric,[7, 8] including millions of individuals who track their steps using wearable accelerometer devices that are available commercially.[9] Despite the longstanding use of step counting in public health interventions,[10] the Physical Activity Guidelines Committee has not yet created recommendations for the number of daily steps to target as a goal for health promotion.[1] The primary reason for this lack of step count guidelines has been a lack of evidence, but meta-analyses conducted from large cohort studies

have recently reported that step count is associated with lower risk of death and chronic disease. [11, 12] Many accelerometers and pedometers have been validated to accurately count steps in the laboratory setting,[13-15] but a remaining concern is that it is unclear how the number of steps reported in studies using research-grade accelerometers compares to steps counted by consumer-grade wearable devices used by the public living in the community (i.e., the free-living setting).

During a recent Framingham Heart Study (FHS) exam cycle, physical activity was measured using both a consumer/mobile health device (Apple Watch) and a research-grade accelerometer (Actical) at the same time in the same individuals. The purpose of this investigation was to assess the agreement between Apple Watch- and Actical-derived daily step count worn in the free-living environments. We primarily assessed whether step count agreed when devices were worn on the same day, even if wear times differed, because we acknowledge that wear time and behavior may differ when participants wear different devices in the real world. We secondarily assessed whether agreement differed when devices were worn for the same hour block and whether agreement differed by age, sex, height, body mass index (BMI), or those with mobility disabilities. This report will enable a better interpretation of the Apple Watch's daily step count for research studies and consumers using these devices.

Methods

Study Cohort: The FHS Third Generation-based (Gen 3) cohort was recruited in 2002-2005 (n=4095),[16] and consisted mostly of grandchildren of the Original FHS cohort,[17] who were largely individuals of European descent. The Gen 3 based examinations also included spouses of the Original cohort’s offspring (New Offspring Spouses, NOS, n=103) who were not already included in the Offspring (Generation 2) cohort and included a multiethnic Omni Group 2 (n=410). Participants from these cohorts have been examined every 6-8 years. All participants provided written informed consent, and the Institutional Review Board at Boston University Medical Center approved the study protocols.

During the third in-person research examination of these cohorts (April 2016 - March 2019), participants were asked to wear an Actical accelerometer for 8 consecutive days on the hip. Beginning in November 2016, as part of the electronic FHS (eFHS) ancillary study,[18] participants also were asked to wear an Apple Watch (Series 0) on their wrist for up to one year if they owned an iPhone with a compatible iOS (version 9 or higher). Of the 3,486 FHS participants examined at the Research Center for exam 3, from April 2016 through March 2019, 2,898 (83%) agreed to take the **Actical monitor**, of which n=2423 (92% of those who took the device) had “valid” steps data, meaning they wore the monitor for at least 3 days, for at least 10 hours/day (**Figure 1**).

In total, 1061 eFHS enrollees (since November 2016) agreed to take the **Apple Watch** or use their own, of which n=959 (90% of those who agreed to use an Apple Watch) wore the device for at least 3 days for at least 10 hours/day during the follow-up period. A total of 834 participants had at least 3 days of “valid” data from both devices (Actical and Apple Watch). Of those, 523 participants had at least 10 hours of wear time on both devices on the same day, providing a total of 3223 person-days for our primary study sample (Sample 1).

Actical Physical Activity

During the 8-day wear period, participants were asked to remove the Actical accelerometer (Philips Respironics, model numbers 198-0302-xx, Respironics Co. Inc, Bend, OR) each night for sleep and when bathing or swimming. Actical data were recorded in 30-second epochs and expressed as counts (or steps) per 30 seconds. Actical step counting has been validated against hand-counting in a laboratory setting.[19, 20] For Sample 1, data were processed using a SAS program developed by Colley et al.,[21] and modified with input from collaborators,[22] including non-wear time removal using the Choi algorithm,[23] as explained in detail in Supplemental Methods. After processing, there remained 18 hours of possible wear time per day. A valid day was defined as ≥10 hours of wear time, with ≥3 days required for inclusion in the main analysis.[24]

Apple Watch Series 0 Physical Activity

As part of the eFHS protocol, participants were asked to wear the smartwatch daily and were sent home with instructions on proper smartwatch use with advice to remove the smartwatch for charging every night. We also set up permissions for the Apple Watch app to access health information from other apps on the smartphone (i.e. steps, heart rate, blood pressure, weight, etc.) but we did not enter participant specific data during Apple Watch set up. In contrast to data collected from the Actical, which had both counts and steps per 30-second interval, we were only able to collect Apple Watch data at the granularity of the number of steps per hour. For the Apple Watch, one wear hour was defined as an hour with at least two heart rates or at least 30 steps accumulated.[25] Unlike for Actical, there was no maximum number of wear hours chosen for the Apple Watch. A valid day was defined as ≥10 hours of wear time, with ≥3 days required for inclusion in the main analysis.

Covariates

The following covariates were measured at the examination that Actical and Apple Watch devices were provided to participants: current smoking status, self-reported health, BMI, hypertension stage II (systolic blood pressure≥140 mmHg or diastolic blood pressure≥90 mmHg and/or use of blood pressure medications),[26] diabetes mellitus (fasting plasma glucose ≥126 mg/dL and/or use of medications for diabetes mellitus), and prevalent cardiovascular disease. Depression status was defined as anyone with a score of 16 or greater on the Center for Epidemiological Studies Depression (CESD) scale. The physical activity index was a composite score constructed by weighting self-reported time spent in physical activity intensities over a 24 hour “typical” day.[27] Mobility limitation was defined as those self-reporting that they were unable to walk 0.5 mile without help or that they were limited a little or a lot when climbing several flights of stairs.

Statistical Analysis

After excluding participants who did not have at least 3 days of valid data from both devices and

then excluding dates on which only one device was worn, we were left with 523 participants (3223 person-days, Sample 1). We compared the number of hours participants wore each device on average days and average steps accumulated to determine device-specific differences, reporting means and standard deviations (SD) or medians and quartile (Q)1 and Q3 in **Table 1**.

To examine agreement between devices on days when both devices were worn for >10 hours (Sample 1), we reported the intraclass correlation (ICC) using the random-effects model in our two study samples and used Lin's concordance coefficient (accounting for repeated observations). We also used kappa coefficients to assess concordance between the devices in identifying participants meeting thresholds of average daily steps (at 3000, 6000, 8000, and 10000 steps/day). Bland-Altman plots were also used to assess potential non-systematic differences between devices and provide a visual representation of these differences in steps/day and the percent differences [100*(Apple Watch mean – Actical steps)/mean steps]. We assessed the limits of agreement for the Bland-Altman plot using repeated measures. Agreement of Apple Watch and Actical step counts per day was also assessed as the percent of days in which steps for each device fell within 15% agreement of one another. In personal communication with physical activity research experts (unpublished), most suggested that acceptable agreement should be set at a 5% difference level (as cited in Tudor-Locke C, et al.[28]), with a few experts acknowledging that agreement within 15% may be considered acceptable (as in Breteler, et al.[14]). Experts polled were those who participated as an author in the meta-analysis of 15 international cohorts with accelerometer data published by Paluch A, et al. 2022.[11] We chose to report the more lenient agreement threshold in order to better detect variability in agreement amongst subsamples of our population, especially after observing the poor overall agreement within these ranges displayed in results.

In secondary analyses, we also examined agreement between devices during hours when both devices were worn (Sample 2) to account for potential differences in wear periods (by device) throughout the day. To create this sample, first, we identified blocks of time ≥3 hours each day

(midnight to midnight) during which both devices were worn. We defined an hour of Actical wear as any hour with >0 step count. In the current study, we defined an hour of Apple Watch wear was defined as an hour with >30 step counts or 0-30 step counts with at least two heart rates recorded, but there does not seem to be an established threshold used in this research field. We excluded hours for which step counts were missing (shown as *NA* in **Supplemental Figure 1**). These definitions differed because of different device wearing locations (hip vs. wrist). When devices are worn on the hip, they can show 0 step counts for prolonged periods of time when a participant is wearing the device sitting, but this is less likely to occur with a wrist-worn device. Thirty participants had <3 hours of overlapping wear time and were excluded (**Figure 1**). These 30 participants had >10 hours Apple Watch wear time on days when the Actical was worn for >10 hours, but the Apple Watch wear hours did not have at least 3 consecutive hours. Each hour or two when steps were counted were broken up by hour(s) with heart rate measurements, but were often missing step counts. An example of 24 hours of Actical and Apple Watch data is shown in **Supplemental Figure 1**. We provide further interpretation of these “interruptions” in wear time in the discussion section. Our next step was to remove the first and last hour of each ≥3-hour block because we could not determine whether they are full or partial hours. The remaining hour(s) in that block were each used as separate data points, to give us steps accumulated by the two devices for every hour that both devices were worn. Because Apple Watch, but not Actical, changes time stamps during the collection period to be consistent as people move across different time zones, we additionally excluded participants residing outside of the Eastern Standard Time Zone (n=36), which may have resulted in discordant hours being counted by each device. One extreme outlier (1 person-hour) was also removed (see **Supplemental Figure 2**), which did not affect results (data not shown). We repeated analysis from Sample 1.

Next, for each sample, we tested for interactions by age, sex, height, and BMI in the linear regression analysis to assess whether these factors influenced agreement between the Apple Watch and Actical device measures of total daily steps. Finally, we performed sensitivity analyses,

repeating our agreement analysis in subsamples excluding participants with high or low step counts. All statistical analyses were performed with R (Version 4.1.3), including the following packages: ggplot2 (for plots), irr (for ICC), epiR (for Lin's concordance correlation), psych (for Kappa coefficients).

Results

Compared to the total FHS Gen 3/NOS/Omni 2 cohort, participants who returned valid (i.e., sufficient) data from the two wearable devices were on average younger, healthier (less smoking, diabetes, hypertension, cardiovascular disease, and depression), and were more likely to have completed college or received a graduate degree (**Table 1**). The average wear time for the Apple Watch was more than an hour longer each day than for the Actical (15.6 vs. 14.4 hours, **Table 1**), which may be partially due to the removal of 6 hours of each 24 hours and other Actical data processing, as described in the **Supplemental Methods**.

Primary Analysis (Sample 1, n=523): Step agreement per day of device wear

We observed modest correlation (ICC=0.56, 95% confidence interval [CI] = 0.54, 0.59, **Table 2**), but poor agreement (29.7% of days having steps counts with ≤15% difference) between devices. Lin's concordance coefficient, accounting for repeated observations, produced the same coefficients as traditional ICC for all results. The two devices demonstrated moderate agreement for distinguishing between participants meeting vs. not meeting step/day thresholds by their average daily steps (kappa coefficient ~0.5, **Table 3**). The Apple Watch and Actical devices were concordant 74.8% - 84.5% of the time, depending on the threshold (3000, 6000, 8000, 10000 steps/day). This reliability for distinguishing between thresholds did not change greatly if we used average daily steps (as in **Table 3**) or steps per person-day (as in **Supplemental Table 1**), but improved slightly to 77.2% - 85.3% if we excluded person-days in which wear time was >1 hour different between devices (**Supplemental Table 1**).

On average, we observed more steps/day counted by the Actical device, with a mean difference of 499 more steps/day counter compared to the Apple Watch (**Figure 2, Table 2**). Limits of agreement were -9543, 8544 steps/day, meaning that differences in step counting between devices is expected to be +/– roughly 9000 steps in a given day of device wear. The differences in step counting between devices tended to increase with higher average steps counted, but percent differences did not (average limits of agreement were -134.6 to 118.2% difference between step counts). There also did not appear to be major under- or over-estimation of steps by one device compared to the other. We observed an interaction (**Supplemental Table 2**, $p<0.0001$) between wear time and device type in their association with daily step count.

Secondary Analysis (Sample 2, n=456): Step agreement per hour of device wear

We conducted secondary analyses to explore the agreement between devices, with differences in wear time minimized. We assessed agreement between devices for each hour during which both devices were worn (456 participants, 1986 person-days, 18760 person-hours, **Table 2** and **Figure 3**). Among hours when both devices were worn, the correlation of absolute step counts between devices was much stronger (ICC=0.86, 95% CI=0.85, 0.86, **Table 2**) than it was for Sample 1, but agreement of steps counted per hour was still poor (only 27.3% of hours having step counts with ≤15% difference) between devices. The mean difference in step count between devices was only 20 steps/hour, but limits of agreement were large (-844, 884 steps/hour) and a 16.6% difference (-98, 131.3% limits of agreement) between Apple Watch and Actical step counting on hours when both devices were worn.

Next, we assessed potential interactions (in Sample 2, **Supplemental Table 2**), observing interactions by obesity status and mobility status ($p<0.001$). We observed that correlations were similar regardless of these factors (Samples 2A-C, **Table 2**), but agreement of step counts with ≤15% difference between devices was slightly worse for participants with obesity (25.9% agreement) or self-reported mobility limitations (23.9% agreement), compared to those with neither

(28.0% agreement).

Sensitivity Analyses: Exploration of days with low step counts and large differences in step count

Despite strong correlation in step counts, there was substantial variability between devices in terms of device agreement, as demonstrated in the total Sample 1, **Figure 2**. We observed 17 person-days with >30,000 steps/day by Actical but <20,000 steps/day by Apple Watch, representing days from 5 participants (**Figure 2, Section A, Supplemental Figure 3, Supplemental Table 3**). In sensitivity analyses excluding data from these 5 participants, the ICC improved slightly for Samples 1 and 2, but the percent of days or hours during which the devices agreed within 15% only improved by <1% (**Supplemental Table 4**). Very few days (~10%) met the more strict 5% threshold for agreement between devices; and agreement was further reduced when only observing days when the Apple Watch was worn 5-10 hours and Actical was worn >10 hours.

Other substantial variability we observed between device counting by Actical compared to Apple Watch was observed in the large number of days during which one device counted <1000 steps and the other device counted >1000 steps (**Figure 2**, Sections B and F). In **Supplemental Figures 4 and 5**, we show scatterplots for hours when both devices were worn. During most hours represented in Sections B and F from **Figure 2**, the devices were either being worn at different times of the day or there were interruptions in step counting. Furthermore, when observing hours from “other days” of those participants who had days that fell into Sections B or F (**Supplemental Figures 4 and 5**), the pattern appears similar to the scatterplot for the overall Sample 2 (**Figure 3**). Exclusion of participants who contributed days that fell into **Figure 2** Sections A, B, D, E or F (those with step counts <1000 or >30000 by either device) did not improve agreement results greatly either, improving the percent of days on which the devices agreed within 15% only up to 32.2% of all days (**Supplemental Table 4).**

Discussion

Consumer accelerometer devices are being used by millions of people to track their physical activity levels and progress toward public health recommendations or personal goals. These devices have been validated in laboratory settings against research-grade devices, but few studies have explored how consumer and research-grade accelerometer step counting compares when participants are living out in the community.

In the current study, we observed poor overall agreement between steps counted by Actical and Apple Watch (series 0) devices. Larger between-device differences were seen when step count was higher, but the percent difference did not increase. However, our results suggest that we can expect the two devices to classify individuals into the same step thresholds about 75-85% of the time. Results such as these may be important to consider when translating future step guidelines to the public using consumer brand devices. The limitations in agreement among accelerometer devices may be less important when they are used by individuals to determine achievement of a recommended number of steps or for the purposes of tracking their step count over time.

The agreement we observed in the current study in a free-living environment was worse than previous laboratory-based studies of consumer-grade devices comparing them to hand-counted steps or research accelerometer devices.[13, 15] However, one study observed that even when testing consistency of step counting in the same device, wearing the device at different locations (wrist vs. hip) can result in inconsistencies in device step counting.[29] The difference in device location alone may have contributed greatly to the poor agreement of step counting between devices in our study.

Similar to our design, one study by Breteler, et al., examined Apple Watch step counting in a free living setting (wrist worn) in comparison to other accelerometers worn on the hip.[14] In this study, 30 healthy participants (mean age 40 years) wearing multiple devices over a 3 day period observed a median absolute relative difference of 7.7% comparing the Apple Watch to the Actigraph (similar

to our mean relative difference results comparing Apple Watch to Actical). However, they did not report the limits of agreement for this relative difference. Other devices they tested had a median absolute relative difference >15%.[14] A low mean or median relative difference indicates low bias (lack of systematic over- or under-counting by one device), but only limits of agreement can inform about the precision of agreement. Breteler, et al. reported the mean difference in step counting was 968 more steps/day counted by the Apple Watch with limits of agreement ±6,000 steps/day (compared to Actigraph), which is almost as high as the limits of agreement we observed in our study sample 1 (compared to Actical). Investigators in that study observed that Apple Watch devices added steps overnight when other devices were not counting any steps, which could have been due to delayed transmission of step count data. We did not observe the same phenomenon in our data, which may be due to us using an older Apple Watch device model. In our study, we observed that some participants had long periods of consecutive hours with heart rate data, but zero step counts (meaning the Apple Watch device was being worn) and had long periods of consecutive Actical step data >0 during this time. We suspect that the Apple Watch step data was either not being recorded/transmitted during these time periods or was delayed by many hours. In order for Apple Watch Series 0 data to be recorded/transmitted, the participant's smartphone needed to be charged, connected to the internet, and unlocked. This finding has important implications for future research teams when analyzing data from other mobile health devices.

Although we observed poor overall agreement (due to wide limits of agreement) in our study and in Breteler, et al.[14] we also reported low bias due to low mean difference and percent difference. However, individual differences in gait, which may be in part due to older age, mobility limitations, or body stature (influenced by sex, height, and body composition), might introduce systematic bias into the measurement of steps in the community and should be considered in future studies.[30-32] Accelerometers have different sensitivities to slow gait speeds or low frequency movement,[29] even when tested in a laboratory in which gait differences are minimized. We observed slightly worse agreement between devices for individuals who were obese or self-reported mobility

limitations. An individual's usual cadence and the amount of time they spend participating in movement activities other than ambulation (such as household chores or other multidirectional activities) may also influence step detection in certain accelerometer devices.[33-35] Although individuals with mobility limitations and other conditions that alter gait (e.g. obesity) only worsened agreement slightly, the overall influence of gait on how these devices count steps may partially explain the poor agreement between devices.

Agreement of step-counting devices has implications for future updates of the Physical Activity Guidelines. Advancements in technology and the widespread availability of consumer wearable devices make physical activity monitoring feasible in the research/clinical settings and for individuals in the community. During the development of the 2018 Physical Activity Guidelines, it was determined that there was insufficient evidence to create a guideline for health promotion based on step count.[1] But an estimation by Tudor-Locke et al. (2011) suggests that the MVPA guidelines can be met by adults who walked a minimum of ~7000-8000 steps/day.[36] Furthermore, a 2022 meta-analysis of 15 observational cohort studies (including FHS) using research-grade physical activity monitors (e.g., Actigraph, Actical, Activ-PAL, etc.), reported that individuals achieving ≥8,000 (vs. <8000) steps/day in middle age had the lowest risk of mortality.[11] In older adults, a lower threshold of ≥6,000 steps/day was associated with almost 50% lower risk of death compared to older adults who walked less. The study, which was the largest meta-analysis of its kind, may serve as evidence to support future guidelines as to the number of steps adults should walk each day.

Although we now have some evidence that achieving step thresholds from 6000-8000 steps/day is associated with lower mortality,[11] creating step guidelines is complicated by the observation that individuals in the community do not use the same research-grade devices as used in many prior studies. Instead, the public uses popular consumer activity trackers, such as Fitbit, Apple Watch, and Garmin, among other devices. Although these consumer devices have been well validated for the measurement of steps in laboratory settings,[13-15] it has not been clear whether the steps

counted by these consumer devices are comparable to steps counted by research-grade devices in free-living settings.[37] Unfortunately, it does not appear to be a simple fix to “convert” steps measured by a research device to those measured by a consumer device, based on the wide limits of agreement. Despite poor overall agreement of step counting between devices, favorably, the devices had substantially better agreement in identifying who meets thresholds between 6000-10000 steps/day, with agreement for ~75-82% of individuals. These thresholds may serve as targets for future public health recommendations.

Strengths and Limitations

Strengths of our investigation include the large sample size and the study being conducted in a community setting, which increases the generalizability of the findings. However, the homogeneous nature of our study cohort, who were mostly from one geographic location, were generally healthier and more highly educated than the general US population, and were mostly of European ancestry, may limit generalizability to more diverse populations. Another strength was our use of different person-day samples to examine different questions such as comparing step counts between devices when worn for a comparable number of hours and observing the influence of different wearing behavior on step count agreement. The lack of control of wear time and differences in wear time observed between the devices can be viewed as both a weakness (because wear time affects step accumulation) and a strength (in that it demonstrates the differences that may be inherent in real-world device use). Similarly, as noted earlier, another difference between these devices was their placement on the wrist versus the hip, which may have also contributed to the variation. However, the device placement locations add another real-world element to our study design.

Wear time per day was longer, on average, for the Apple Watch, which may, in part, be due to wearing during sleeping hours. In our analysis, sleeping hours were removed from the Actical total wear time, but not from the Apple Watch. We asked participants to remove the Actical device when they bathed, swam, or slept. These instructions were not given to participants for the Apple Watch,

although participants may have chosen to do so. The Apple Watch is waterproof, but the battery does not typically last much longer than 24 hours, so most participants likely took off the Apple Watch to charge at night. If the Apple Watch battery was not charged, a participant might not have worn the device and may have missed opportunities to record steps walked. The Actical battery did not need to be charged during the week that participants wore the device, which may have affected when it was worn compared to the Apple Watch. On the other hand, the Actical device was worn on a belt around the waist/hip, compared to the Apple Watch, worn on the wrist, either of which can be cumbersome, causing some participants to remove the device or wear it improperly (e.g., loosely). It is unclear which placement site is preferred by the research community.[38] Although we sent participants home with instructions for when to take on and off the devices and the location they should be worn, we did not emphasize that they should ensure a snug fit. Another possibility is that there could be calibration issues with some of the devices (Actical and/or Apple Watch could have drifted from factory calibration). Comparison of agreement results between Sample 1 and 2 make it clear that it is unlikely that the poor agreement was explained by participants wearing the devices during different times of day or activities. But it is also evident that some of the differences in step counting by these devices may have been due to Bluetooth connectivity errors in the recording or transmission of step data, that led to very low steps counted by the Apple Watch.

Features of physical activity monitors are also important considerations. The Apple Watch device used in this investigation has many applications, including allowing participants to see step counts as they were accumulated (there was no visual display on the Actical device) and other functionalities. The availability of these features may also influence when the device was worn and how many steps are taken. We did observe that of the participants who agreed to either device, a roughly equal proportion of participants (~90%) wore those devices for ≥3 days for at least 10 hours/ day. But studies have shown that features such as a display showing step progress and encouragement (i.e., nudges) to stand or move may increase both wearing and stepping behavior, especially over the short term, which may influence results from studies using consumer devices

that tend to have these features.

Our study provided us with many lessons that we hope to communicate with investigators using accelerometers. An unexpected finding was that the agreement between these physical activity monitors was only improved slightly after we limited differences in wear time between the devices. When experts develop public health guidelines for the number of steps to walk each day, they must consider that devices do not all record steps equivalently and that the type of device, wearing location, or mode (i.e., watch, belt, or smartphone app), battery life, Bluetooth connectivity issues, other features of the device, and gait differences of participants may all influence when the device is worn and how many steps are counted. Moreover, we did not enter participant-specific data (i.e. height, weight, age, sex) when setting up the Apple Watch or Actical devices. However, the Apple Watch may have accessed this type of data from other health-related apps on a participant's smartphone. It is also important to note that we studied older versions of the devices, both of which are no longer supported by their manufacturers. Hopefully, newer device models may have overcome some of the limitations of the accelerometers we studied; we used Apple Watch Series 0 during data collection, but they have already transitioned to Series 8. In future research, it may also be important to emphasize proper wear of devices and input relevant participant-specific information during device set up for improved precision.

Conclusion

Our investigation suggests that overall agreement between steps counted by the Actical and Apple Watch Series 0 devices was poor, but agreement between devices was much stronger for distinguishing who meets certain step thresholds. Many large cohort studies have used the Actical device and other research- and consumer-devices to observe thresholds of physical activity (steps per day) that are associated with health outcomes.[11, 39-41] Lessons learned from our investigation should be considered when translating thresholds of steps counted using the Actical to guidelines for members of the community using consumer devices, including the Apple Watch.

Future studies should explore the agreement among other devices in the community setting and explore the role of interruptions in connectivity, calibration, and factors affecting gait, such as age, sex, frailty/mobility status, BMI, and height on the accuracy of step count and agreement among devices.. However, another important future direction should be the increased use of consumer accelerometer devices in research in order to replicate recent meta-analyses reporting the higher risk of mortality among physically inactive individuals (measured using research-grade devices).[11, 42] Studies such as All of Us and the RURAL Heart and Lung Study that use Fitbit devices, for example, will be extremely useful in the development and translation of future physical activity step guidelines.[41]

The good news is that the impact of these challenges in measuring steps may be minimized when accelerometers are used by individuals for the purposes of tracking the changes in their physical activity over time, which eliminates the impact of gait differences (unless gait changes), factory calibration issues (if the same device is used), and presumably connectivity issues would remain consistent, limiting their impact too.

Acknowledgments: This study was supported by an award from the Robert Wood Johnson Foundation (number 74624) and grants from the National Heart Lung and Blood Institute (R01HL141434, R01HL131029), National Institute on Aging (R01AG047645), and American Heart Association (15GPSGC24800006). The Framingham Heart Study was supported by a contract from the National Heart Lung and Blood Institute (principal investigator RSV 75N92019D00031); and investigator time from the following grants: R01HL126911 (EJB), R01HL092577 (EJB), R01AG066010 (EJB), U54HL120163 (EJB), R01HL155343 (DDM), R01HL141434 (DDM), R33HL158541 (DDM), U54HL143541 and U54HL143541-05S1, UG3NS135168 (DDM), American Heart Association, 18SFRN34110082 (EJB). Dr. Vasan was supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University Chobanian and Avedisian School of Medicine. The Apple Watches were provided to

Boston University by Apple Inc at no cost to the study. The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

Conflicts of Interest: Apple was not involved in the study design, analysis, interpretation, or reporting of study results. DDM has received research support from Fitbit, Apple Inc, Bristol–Myers Squibb, Boehringer–Ingelheim, Pfizer, Flexcon, Samsung, Philips Healthcare, and Biotronik, and he has received consultancy fees from Heart Rhythm Society, Bristol–Myers Squibb, Pfizer, Fitbit, Flexcon, Boston Biomedical Associates, VentureWell, Avania, NAMSA and Rose Consulting. DDM also declares financial support for serving on the Steering Committee for the GUARD-AF study (NCT04126486) and the advisory committee for the Fitbit Heart study (NCT04176926). VK is a principal, and CN is an employee of CareEvolution, Inc, a healthcare technology company. NLS received funding from Novo Nordisk for an investigator-initiated research grant unrelated to the current paper. JMM received funding as a guest lecturer for Merck unrelated to this work. The remaining authors declare no conflicts of interest.

References

[1] 2018 Physical Activity Guidelines Advisory Committee (2018) 2018 Physical Activity Guidelines Advisory Committee Scientific Report. *Washington, DC: U.S. Department of Health and Human Services*.

[2] Tucker JM, Welk GJ, Beyler NK (2011) Physical activity in U.S.: adults compliance with the Physical Activity Guidelines for Americans. *Am J Prev Med* **40**, 454-461.

[3] Zenko Z, Willis EA, White DA (2019) Proportion of Adults Meeting the 2018 Physical Activity Guidelines for Americans According to Accelerometers. *Front Public Health* **7**, 135.

[4] Lorbergs AL, Prorok JC, Holroyd-Leduc J, Bouchard DR, Giguere A, Gramlich L, Keller H, Tang A, Racey M, Ali MU, Fitzpatrick-Lewis D, Sherifali D, Kim P, Muscedere J (2022) Nutrition and Physical Activity Clinical Practice Guidelines for Older Adults Living with Frailty. *J Frailty Aging* **11**, 3-11.

[5] Fanning J, Nicklas BJ, Rejeski WJ (2022) Intervening on physical activity and sedentary behavior in older adults. *Exp Gerontol* **157**, 111634.

[6] Office of the Surgeon G (2015) Publications and Reports of the Surgeon General In *Step It Up! The Surgeon General's Call to Action to Promote Walking and Walkable Communities* US Department of Health and Human Services, Washington (DC).

[7] Fox S, Duggan M (2013) Tracking for Health. *Pew Internet and American Life Project. Pew Research Center* **Accessed June 27, 2023.** **http://www.pewinternet.org/2013/01/28/tracking-for-health/**.

[8] Abril EP (2016) Tracking Myself: Assessing the Contribution of Mobile Technologies for Self-Trackers of Weight, Diet, or Exercise. *J Health Commun* **21**, 638-646.

[9] Deloitte (2022) Connectivity and Mobile Trends Survey, 2nd edition. **Accessed June 27, 2023.** **https://www2.deloitte.com/us/en/insights/industry/telecommunications/connectivity-mobile-trends-survey.html**.

[10] Bassett DR, Jr., Toth LP, LaMunion SR, Crouter SE (2017) Step Counting: A Review of Measurement Considerations and Health-Related Applications. *Sports Med* **47**, 1303-1315.

[11] Paluch AE, Bajpai S, Bassett DR, Carnethon MR, Ekelund U, Evenson KR, Galuska DA, Jefferis BJ, Kraus WE, Lee IM, Matthews CE, Omura JD, Patel AV, Pieper CF, Rees-Punia E, Dallmeier D, Klenk J, Whincup PH, Dooley EE, Pettee Gabriel K, Palta P, Pompeii LA, Chernofsky A, Larson MG, Vasan RS, Spartano N, Ballin M, Nordström P, Nordström A, Anderssen SA, Hansen BH, Cochrane JA, Dwyer T, Wang J, Ferrucci L, Liu F, Schrack J, Urbanek J, Saint-Maurice PF, Yamamoto N, Yoshitake Y, Newton RL, Jr., Yang S, Shiroma EJ, Fulton JE (2022) Daily steps and all-cause mortality: a meta-analysis of 15 international cohorts. *Lancet Public Health* **7**, e219-e228.

[12] Hall KS, Hyde ET, Bassett DR, Carlson SA, Carnethon MR, Ekelund U, Evenson KR, Galuska DA, Kraus WE, Lee IM, Matthews CE, Omura JD, Paluch AE, Thomas WI, Fulton JE (2020) Systematic review of the prospective association of daily step counts with risk of mortality, cardiovascular disease, and dysglycemia. *Int J Behav Nutr Phys Act* **17**, 78.

[13] El-Amrawy F, Nounou MI (2015) Are Currently Available Wearable Devices for Activity Tracking and Heart Rate Monitoring Accurate, Precise, and Medically Beneficial? *Healthc Inform Res* **21**, 315-320.

[14] Breteler MJ, Janssen JH, Spiering W, Kalkman CJ, van Solinge WW, Dohmen DA (2019) Measuring Free-Living Physical Activity With Three Commercially Available Activity Monitors for Telemonitoring Purposes: Validation Study. *JMIR Form Res* **3**, e11489.

[15] Xie J, Wen D, Liang L, Jia Y, Gao L, Lei J (2018) Evaluating the Validity of Current Mainstream

Wearable Devices in Fitness Tracking Under Various Physical Activities: Comparative Study. *JMIR Mhealth Uhealth* **6**, e94.

[16] Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, D'Agostino RB, Sr., Fox CS, Larson MG, Murabito JM, O'Donnell CJ, Vasan RS, Wolf PA, Levy D (2007) The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am J Epidemiol* **165**, 1328-1335.

[17] Dawber TR, Kannel WB (1966) The Framingham study. An epidemiological approach to coronary heart disease. *Circulation* **34**, 553-555.

[18] McManus DD, Trinquart L, Benjamin EJ, Manders ES, Fusco K, Jung LS, Spartano NL, Kheterpal V, Nowak C, Sardana M, Murabito JM (2019) Design and Preliminary Findings From a New Electronic Cohort Embedded in the Framingham Heart Study. *J Med Internet Res* **21**, e12143.

[19] Esliger DW, Probert A, Connor Gorber S, Bryan S, Laviolette M, Tremblay MS (2007) Validity of the Actical accelerometer step-count function. *Med Sci Sports Exerc* **39**, 1200-1204.

[20] Johnson M, Meltz K, Hart K, Schmudlach M, Clarkson L, Borman K (2015) Validity of the Actical activity monitor for assessing steps and energy expenditure during walking. *J Sports Sci* **33**, 769-776.

[21] Colley RC, Tremblay MS (2011) Moderate and vigorous physical activity intensity cut-points for the Actical accelerometer. *J Sports Sci* **29**, 783-789.

[22] Evenson KR, Sotres-Alvarez D, Deng YU, Marshall SJ, Isasi CR, Esliger DW, Davis S (2015) Accelerometer adherence and performance in a cohort study of US Hispanic adults. *Med Sci Sports Exerc* **47**, 725-734.

[23] Choi L, Liu Z, Matthews CE, Buchowski MS (2011) Validation of accelerometer wear and nonwear time classification algorithm. *Med Sci Sports Exerc* **43**, 357-364.

[24] Hart TL, Swartz AM, Cashin SE, Strath SJ (2011) How many days of monitoring predict physical activity and sedentary behaviour in older adults? *Int J Behav Nutr Phys Act* **8**, 62.

[25] Lin H, Sardana M, Zhang Y, Liu C, Trinquart L, Benjamin EJ, Manders ES, Fusco K, Kornej J, Hammond MM, Spartano NL, Pathiravasan CH, Kheterpal V, Nowak C, Borrelli B, Murabito JM, McManus DD (2020) Association of Habitual Physical Activity With Cardiovascular Disease Risk. *Circ Res* **127**, 1253-1260.

[26] Whelton PK, Carey RM, Aronow WS, Casey DE, Jr., Collins KJ, Dennison Himmelfarb C, DePalma SM, Gidding S, Jamerson KA, Jones DW, MacLaughlin EJ, Muntner P, Ovbiagele B, Smith SC, Jr., Spencer CC, Stafford RS, Taler SJ, Thomas RJ, Williams KA, Sr., Williamson JD, Wright JT, Jr. (2018) 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **138**, e426-e483.

[27] Kannel WB, Belanger A, D'Agostino R, Israel I (1986) Physical activity and physical demand on the job and risk of cardiovascular disease and death: the Framingham Study. *Am Heart J* **112**, 820-825.

[28] Tudor-Locke C, Sisson SB, Lee SM, Craig CL, Plotnikoff RC, Bauman A (2006) Evaluation of quality of commercial pedometers. *Can J Public Health* **97 Suppl 1**, S10-15, s10-16.

[29] Mora-Gonzalez J, Gould ZR, Moore CC, Aguiar EJ, Ducharme SW, Schuna JM, Jr., Barreira TV, Staudenmayer J, McAvoy CR, Boikova M, Miller TA, Tudor-Locke C (2022) A catalog of validity indices for step counting wearable technologies during treadmill walking: the CADENCE-adults study. *Int J Behav Nutr Phys Act* **19**, 117.

[30] Tyo BM, Fitzhugh EC, Bassett DR, Jr., John D, Feito Y, Thompson DL (2011) Effects of body mass index and step rate on pedometer error in a free-living environment. *Med Sci Sports Exerc* **43**, 350-356.

[31] Pomeroy J, Brage S, Curtis JM, Swan PD, Knowler WC, Franks PW (2011) Between-monitor differences in step counts are related to body size: implications for objective physical activity measurement. *PLoS One* **6**, e18942.

[32] Feito Y, Bassett DR, Thompson DL, Tyo BM (2012) Effects of body mass index on step count accuracy of physical activity monitors. *J Phys Act Health* **9**, 594-600.

[33] Hickey A, John D, Sasaki JE, Mavilia M, Freedson P (2016) Validity of Activity Monitor Step Detection Is Related to Movement Patterns. *J Phys Act Health* **13**, 145-153.

[34] Dall PM, McCrorie PR, Granat MH, Stansfield BW (2013) Step accumulation per minute epoch is not the same as cadence for free-living adults. *Med Sci Sports Exerc* **45**, 1995-2001.

[35] Fokkema T, Kooiman TJ, Krijnen WP, CP VDS, M DEG (2017) Reliability and Validity of Ten Consumer Activity Trackers Depend on Walking Speed. *Med Sci Sports Exerc* **49**, 793-800.

[36] Tudor-Locke C, Craig CL, Brown WJ, Clemes SA, De Cocker K, Giles-Corti B, Hatano Y, Inoue S, Matsudo SM, Mutrie N, Oppert JM, Rowe DA, Schmidt MD, Schofield GM, Spence JC, Teixeira PJ, Tully MA, Blair SN (2011) How many steps/day are enough? For adults. *Int J Behav Nutr Phys Act* **8**, 79.

[37] Middelweerd A, HP VDP, A VANH, Twisk JWR, Brug J, Te Velde SJ (2017) A Validation Study of the Fitbit One in Daily Life Using Different Time Intervals. *Med Sci Sports Exerc* **49**, 1270-1279.

[38] Schrack JA, Cooper R, Koster A, Shiroma EJ, Murabito JM, Rejeski WJ, Ferrucci L, Harris TB (2016) Assessing Daily Physical Activity in Older Adults: Unraveling the Complexity of Monitors, Measures, and Methods. *J Gerontol A Biol Sci Med Sci* **71**, 1039-1048.

[39] Cuthbertson CC, Moore CC, Sotres-Alvarez D, Heiss G, Isasi CR, Mossavar-Rahmani Y, Carlson JA, Gallo LC, Llabre MM, Garcia-Bedoya OL, Farelo DG, Evenson KR (2022) Associations of steps per day and step intensity with the risk of diabetes: the Hispanic Community Health Study / Study of Latinos (HCHS/SOL). *Int J Behav Nutr Phys Act* **19**, 46.

[40] Spartano NL, Demissie S, Himali JJ, Dukes KA, Murabito JM, Vasan RS, Beiser AS, Seshadri S (2019) Accelerometer-determined physical activity and cognitive function in middle-aged and older adults from two generations of the Framingham Heart Study. *Alzheimers Dement (N Y)* **5**, 618-626.

[41] Master H, Annis J, Huang S, Beckman JA, Ratsimbazafy F, Marginean K, Carroll R, Natarajan K, Harrell FE, Roden DM, Harris P, Brittain EL (2022) Association of step counts over time with the risk of chronic disease in the All of Us Research Program. *Nat Med* **28**, 2301-2308.

[42] Ekelund U, Tarp J, Steene-Johannessen J, Hansen BH, Jefferis B, Fagerland MW, Whincup P, Diaz KM, Hooker SP, Chernofsky A, Larson MG, Spartano N, Vasan RS, Dohrn IM, Hagströmer M, Edwardson C, Yates T, Shiroma E, Anderssen SA, Lee IM (2019) Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: systematic review and harmonised meta-analysis. *Bmj* **366**, l4570.

Figure Legends

Figure 1. Participant flow diagram for the analysis of agreement between Apple Watch and Actical step counts

Abbreviations: Framingham Heart Study (FHS); electronic (e)FHS; Generation (Gen); New Offspring Spouse (NOS); eastern standard time (EST).

*Enrollment in eFHS starting in November 2016 was necessary because this was the first date Apple Watches were given out at the FHS Research Center. Participants were able to enroll in eFHS prior to this, but they were given an Apple Watch to use later (starting in November), so their Apple Watch use would not align with the Actical monitor wearing dates.

Figure 2. Scatterplot and Bland Altman plots (difference and %difference) of Apple Watch steps by Actical steps accumulated on the same date (Sample 1, all data, 3223 person-days, 523 participants).

Each point represents data from one participant on a single date (1 person-day). In the scatterplot, dashed lines are set at 1,000 and 30,000 step thresholds. Days on which participants accumulated 1,000-30,000 steps are dark green, and days outside that threshold are presented in light green. Sections separated by the dashed lines include the following number of person-days/participants: A=17/5, B=205/68, C=2963/512, D=3/2, E=4/4, F=31/29. The Bland Altman plots on the right show the mean difference or mean % difference (red dashed line) and the limits of agreement 95% confidence interval (blue dashed lines). Mean % difference was calculated as [100*(Apple Watch steps minus Actical steps)/ (average Apple Watch and Actical steps)].

Figure 3. Scatterplot and Bland Altman plot (difference and %difference) of Apple Watch steps by Actical steps accumulated during hours when both devices were worn (Sample 2, n=456 participants; n=1986 person-days; n=18760 person-hours).

Each point represents data from a single hour (1 person-hour). The Bland Altman plots on the right

show the mean difference or mean % difference (red dashed line) and the limits of agreement 95% confidence interval (blue dashed lines). Mean % difference was calculated as [100*(Apple Watch steps minus Actical steps)/ (average Apple Watch and Actical steps)].

Supplemental Digital Content: Supplemental Digital Content 1.doc

Table 1. Characteristics for all Framingham Heart Study (FHS) Gen 3 participants who attended exam 3, compared to those with valid Actical and Apple Watch data on the same date

	FHS Gen 3 (n=3521)	FHS Gen 3 with valid Actical + Apple Watch data on the same date (Sample 1, n=523)
Age, years	54.5 (9.4)	51.7 (8.9)
Women (%)	1896 (53.9%)	298 (57.0%)
Race and ethnicity		
Non-Hispanic White	3233 (91.8%)	478 (91.4%)
Non-Hispanic Black	59 (1.7%)	12 (2.3%)
Hispanic/Latino	106 (3.0%)	14 (2.7%)
Asian	71 (2.0%)	9 (1.7%)
American Indian	1 (0.03%)	1 (0.2%)
Pacific Islander	2 (0.06%)	0 (0%)
More than one race	41 (1.2%)	8 (1.5%)
Unknown	8 (0.2%)	1 (0.2%)
BMI, kg/m^2	28.6 (6.2)	28.2 (5.7)
Height, inches	66.6 (3.7)	66.8 (3.6)
Mobility limitation (%)	703 (20%)	85 (16.3%)
Smoking (%)	234 (6.7%)	27 (5.2%)
Education		
Less than HS	48 (1.4%)	3 (0.6%)
Completed HS	470 (13.5%)	43 (8.2%)
Some College	489 (14%)	114 (21.8%)
Bachelor's Degree	1222 (35.0%)	214 (41%)
Graduate/Professional Degree	843 (24.1%)	148 (28.4%)
Married, living as married, living with partner (%)	2454 (70.5%)	397 (76.4%)
Employed full time (%)	2277 (65.4%)	381 (73.1%)
Self-reported health, excellent (%)	750 (21.4%)	128 (24.5%)
Diabetes mellitus (%)	310 (8.8%)	26 (5.0%)
Hypertension stage II (%)	1095 (31.1%)	112 (21.4%)
Cardiovascular disease (%)	164 (4.7%)	18 (3.4%)
Depression [CESD>16] (%)	449 (12.8%)	55 (10.5%)
Physical activity index, score	33.9 (5.7)	33.2 (4.7)
Actical steps, median [Q1, Q3]	--	7064 [4638, 10529]
Apple Watch steps, median [Q1, Q3]	--	7060 [4450, 10348]
Actical wear time, hours	--	14.4 (1.8)
Apple Watch wear time, hours	--	15.6 (2.6)

Data are depicted as n (percent), mean (SD), or median [Q1, Q3]
Abbreviations: BMI, body mass index; CESD, Center for Epidemiological Studies Depression; HS, high school; Q, quarter; SD, standard deviation

Table 2. Agreement between steps accumulated on Actical vs. Apple Watch device by participants wearing both devices on the same date

Sample	Sample description	Adjusted Linear Regression, Beta (95%CI)	ICC (95% CI)	Lin's concordance correlation	Mean difference * (Bland Altman Limits of Agreement)	Mean % difference** (Bland Altman Limits of Agreement)	Percent of Apple Watch days with a step count within 15% agreement compared to Actical
Sample 1 (n=523 participants; n=3223 person-days)	Includes all days when both devices were worn for >10 hours	0.67 (0.65, 0.70)	0.56 (0.54, 0.59)	0.56 (0.54, 0.58)	-499 (-9543, 8544)	-8.2% (-134.6, 118.2)	29.7%
Sample 2 (n=456 participants; n=1986 person-days; n=18760 person-hours)	Only includes blocks of hours during which both devices were worn ***	0.97 (0.96, 0.97)	0.86 (0.85, 0.86)	0.86 (0.85, 0.86)	20 (-844, 884)	16.6% (-98.0, 131.3)	27.3%
Sample 2A – *with obesity* (n=151 participants; n=5397 person hours)		0.94 (0.92, 0.95)	0.85 (0.84, 0.86)	0.85 (0.84, 0.85)	33 (-844, 909)	18.2% (-95.5, 131.9)	25.9%
Sample 2B - *with mobility limits* (n=79 participants; n=2967 person-hours)		0.86 (0.84, 0.88)	0.86 (0.83, 0.88)	0.86 (0.85, 0.87)	98 (-953, 1148)	29.8% (-83.8, 143.5)	23.9%
Sample 2C - *without obesity or mobility limitations* (n=266 participants; n=11699 person hours)		0.98 (0.97, 0.99)	0.85 (0.85, 0.86)	0.85 (0.85, 0.86)	3 (-1089, 1096)	14.4% (-101.4, 130.3)	28.0%

The adjusted linear regression model includes: age, sex, cohort type, BMI, height, (and the difference in wear time for Sample 1)

* Mean difference was Apple Watch steps minus Actical steps

** Mean % difference was [100*(Apple Watch steps minus Actical steps)/ (average Apple Watch and Actical steps)]

***Sample 2: After removing hours when both devices were not being worn, we removed the first and last hours of remaining blocks of hours. We additionally excluded participants who lived outside EST time zone and removed one datapoint that was an extreme outlier (See Supplemental Figure 1). We used each remaining hour as a separate data point.

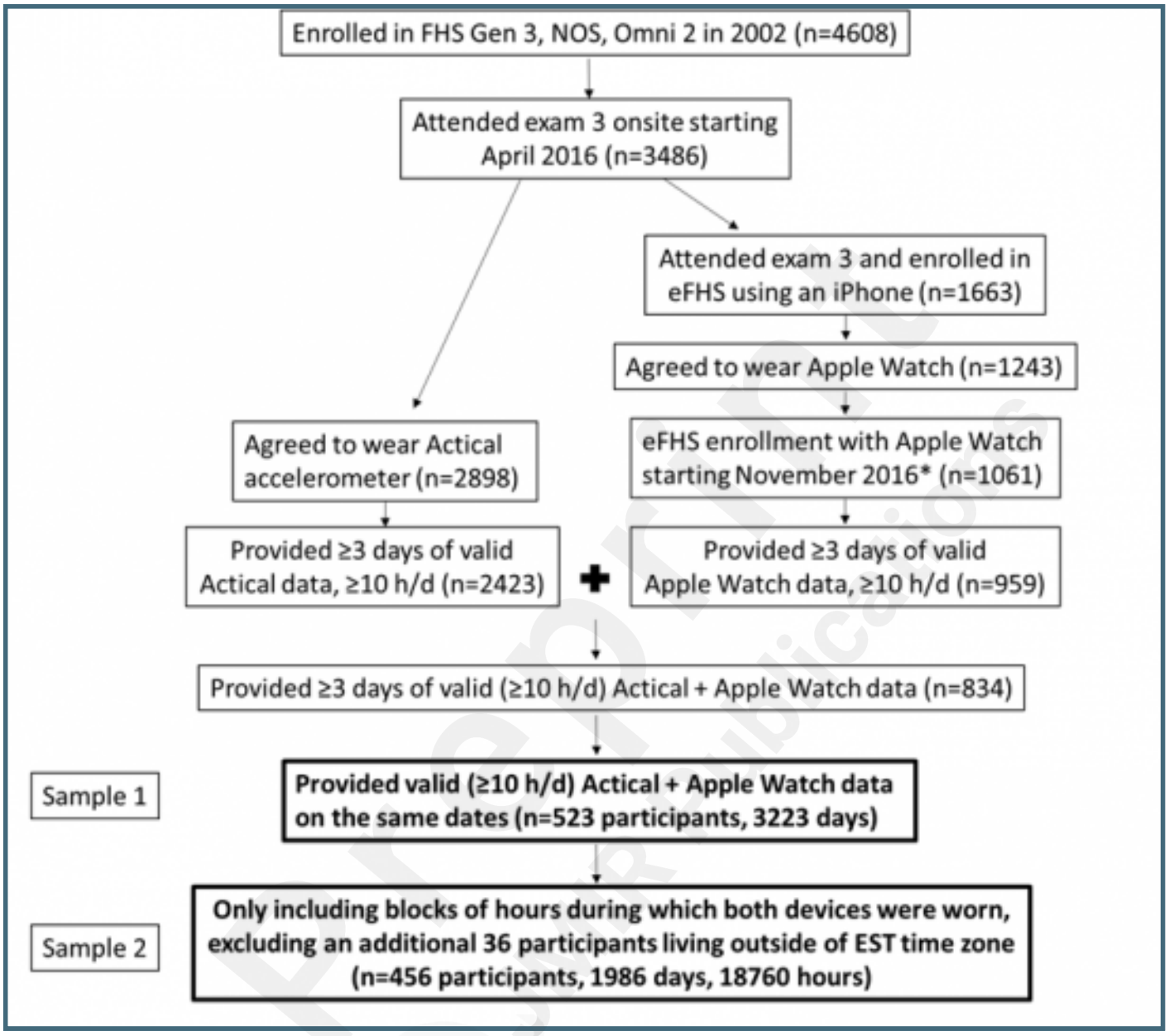
Table 3. Agreement of Actical and Apple Watch devices to identify participants meeting average daily step thresholds (**Sample 1**, n=523 participants; 3223 person-days)

Step/day Threshold	**Percent concordance** for "meets the PA threshold" as measured by the two devices	**Kappa coefficients (95% CI)** for "meets the PA threshold" as measured by the two devices
3000 steps/day	442 (84.5%)	0.12 (0.01, 0.22)
6000 steps/day	396 (75.7%)	0.46 (0.38, 0.54)
8000 steps/day	391 (74.8%)	0.49 (0.41, 0.56)
10000 steps/day	426 (81.5%)	0.49 (0.40, 0.58)
Abbreviation: PA, physical activity		

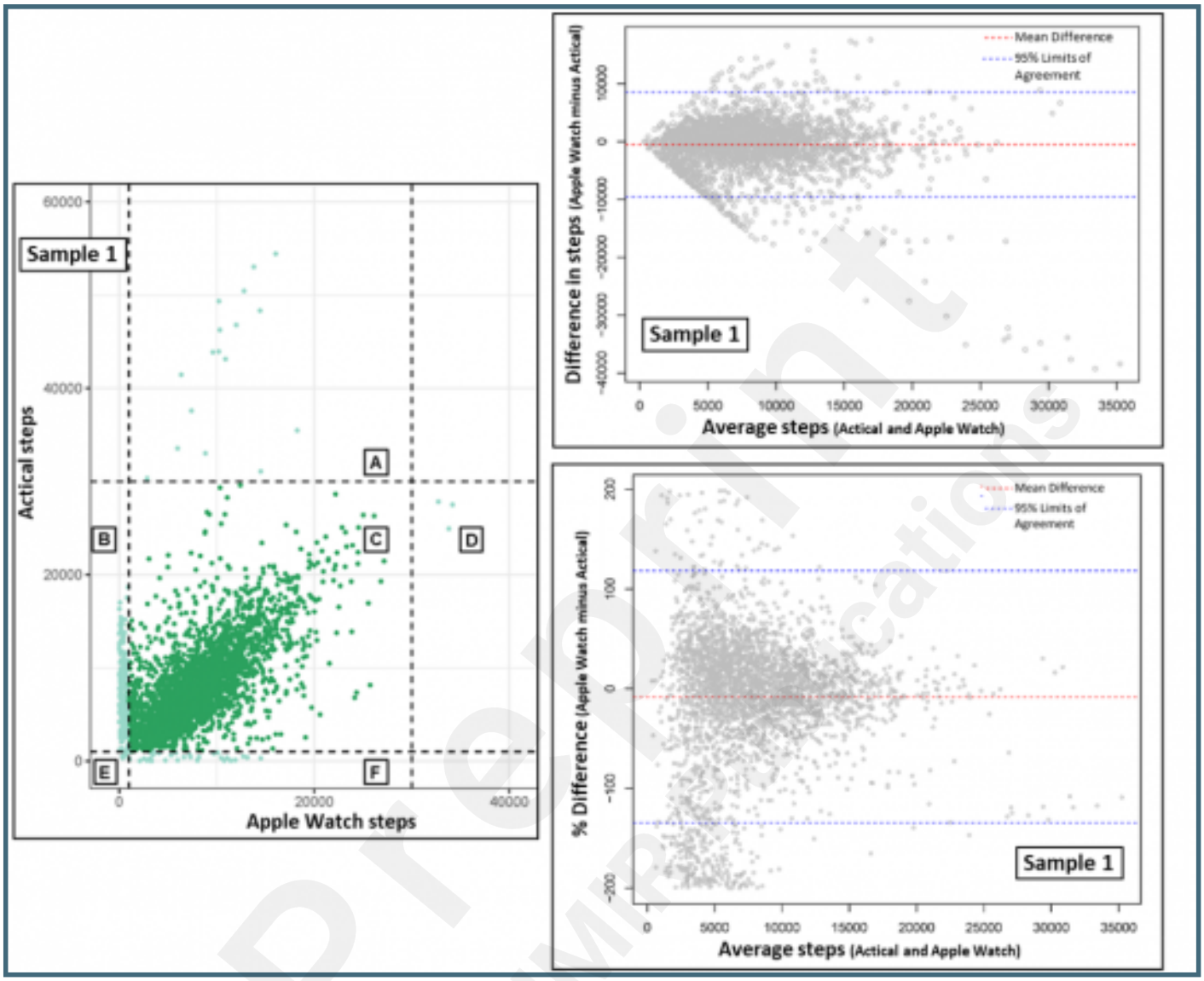
Supplementary Files

Figures

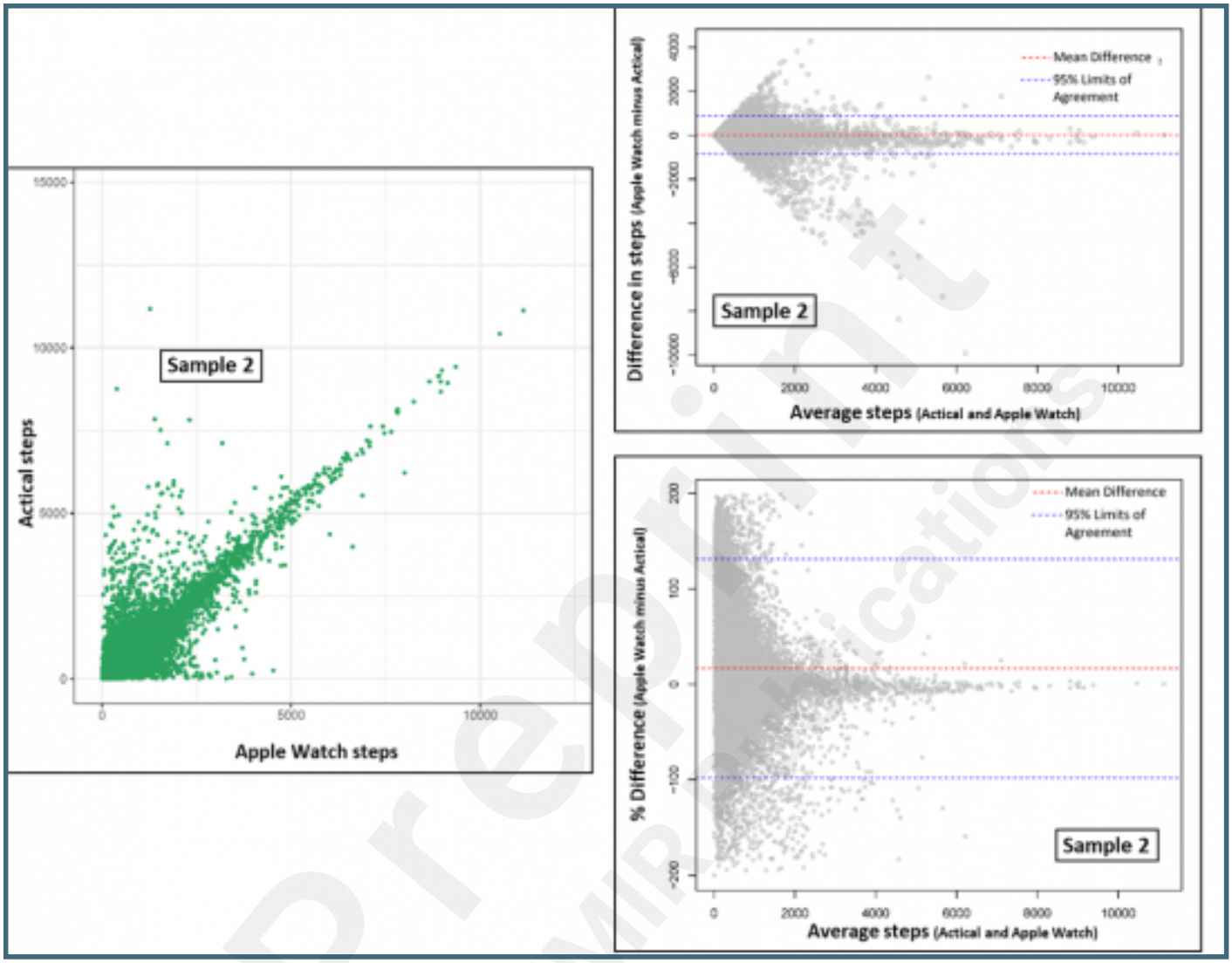
Participant flow diagram for the analysis of agreement between Apple Watch and Actical step counts.



Scatterplot and Bland Altman plots (difference and %difference) of Apple Watch steps by Actical steps accumulated on the same date (Sample 1, all data, 3223 person-days, 523 participants).



Scatterplot and Bland Altman plot (difference and %difference) of Apple Watch steps by Actical steps accumulated during hours when both devices were worn (Sample 2, n=456 participants; n=1986 person-days; n=18760 person-hours).



Multimedia Appendixes

Supplemental Materials.
URL: http://asset.jmir.pub/assets/aa6153f2e1276bd4f5897109d5c8179a.docx