Apple Watch Heart Rate Accuracy Study

September 2026

Overview

Heart rate tracking is a central part of health and fitness wearables. It's used for daily health and workout insights as well as an input to core features, such as calorie estimates, VO_2 max estimates, period predictions, high and low heart rate alerts, and many others. The accuracy of heart rate tracking is therefore critical, and the foundation for meaningful insights in the user experience.

With the release of Apple Watch Series 12 and Apple Watch Ultra 4, Apple introduced the Health Sensing System, which includes both updated electrical and optical (photoplethysmography, or PPG) heart sensors and associated signal processing algorithms, as well as several features enabled by the new technology, such as Recovery HRV and the Readiness app.

This paper summarizes the accuracy of the updated optical heart rate sensing system relative to other devices in the wearables category through a large multi-site head-to-head study, using an ECG chest strap as reference, which is an established standard used for evaluating heart rate accuracy of wearables[1,2]. The study, with over 1,000 participants, evaluates a broad set of comparative devices, across a wide range of activities, with sufficient sample size to allow for statistically powered conclusions.

Study Design

Participants wore an Apple Watch on one wrist, and a comparative wrist-based device on the other, while using a Polar H10 ECG based chest strap as reference. Wrist assignment was randomized such that Apple Watch was worn on the dominant hand approximately 50% of the time. Wrist-based devices studied included Apple Watch Series 12, Garmin Forerunner 970, Google Pixel Watch 4, Huawei Watch 5, Samsung Galaxy Watch 8, and WHOOP 5. Oura Ring 5 was also worn on an index finger, per the manufacturer's stated best practices[3]. These brands represent approximately 90% of the worldwide smartwatch category, as well as two prominent screenless wearable devices[4]. To ensure appropriate fit, moderators helped secure each device such that it was snug, but comfortable and consistent with the manufacturer's recommendation. The latest device model available from each manufacturer was used for evaluation during July and August of 2026, when the study was conducted.

To ensure an adequate range of activities were conducted for generalizable conclusions, participants completed outdoor runs at various speeds, indoor interval treadmill runs, outdoor cycling, high intensity interval training (HIIT), and strength training, with duration ranging from 15 to 20 minutes per exercise type. This mix of activities was specifically selected to include challenging conditions for optical heart rate performance — e.g., cadence lock associated with running, loaded forearm grip associated with strength training and cycling, rapid wrist motion and heart rate transitions associated with HIIT. Participants also completed a Daily Living protocol to evaluate background heart rate accuracy across activities like rest, desk work, meal preparation, and walking, each for approximately 20 minutes. Participants were permitted to complete both the Workout and the Daily Living protocols. The Oura Ring was not evaluated in strength training per the manufacturer's recommendation to avoid wearing the ring during weightlifting[5]. All participants agreed to terms outlined in an informed consent form (ICF). This ICF and protocol were approved by multiple review boards within Apple, which included representation from groups focused on research ethics, safety, privacy, and data governance.

The study was conducted at five sites: two near Apple's headquarters in Cupertino, California, plus sites in San Diego, California; Austin, Texas; and Selangor, Malaysia. Apple managed recruitment and execution at the Cupertino and Austin sites, while a third-party contract research organization (CRO) managed the San Diego and Selangor sites.

The study focused on a diverse population with representation across sex, age, BMI, and skin tone. Inclusion criteria included being at least 18 years of age, able to wear wrist, finger, and chest devices, and able to tolerate physical activity for 30 minutes, and other criteria to ensure participant safety. Exclusion criteria included working in tech media, having a known allergy to study equipment, taking beta blockers, having any contraindication to intense exercise — including unstable cardiovascular conditions. Individuals were also excluded with device-site tattoos or large moles, which may interfere with readings across manufacturers.

In selecting the data stream and source evaluated for each device, we prioritized three primary criteria. First, the data stream needed to be visible to the user in the product experience, as the purpose of the study is to evaluate accuracy experienced by the user. Second, the data stream needed to include appropriate timestamping to support paired analysis with the reference chest strap. Third, when multiple frequency options were available, the highest frequency source was selected to reduce ambiguity and impact of how each manufacturer may have chosen to introduce or omit any type of averaging.

Devices required different methods to obtain heart rate data. We selected a method for each that met all the criteria outlined above. The table below outlines the method and heart rate reporting interval for each device. A Bluetooth Low Energy data stream (BLE) sent from the manufacturer's heart rate app to a data logging app on a separate iPhone was used in many instances because it uniquely met all criteria outlined above, as not all manufacturers write high-fidelity data to HealthKit, or permit such data to be exported from their companion apps. If a device supported a workout mode, that was turned on for workout activities and left off for the Daily Living activities.

To obtain appropriate data from Apple Watch, in addition to data exported from HealthKit, an internal data logging application was used to acquire the highest-frequency data published to the user. This data is a single value (not an average) reported every 5 seconds — in watch face complications for background heart rate, and in the Workout app for workout activities. The internal app, which logs the same data surfaced to user, was required because background data is only published to HealthKit every 30 seconds to manage device memory.

While the reporting interval varied across the comparison set, each user facing data stream was evaluated versus a consistently structured reference stream to appropriately assess the accuracy of the experience surfaced to the user.

Mode	Device	Collection Method	Reporting Interval
Workout Mode	Apple Watch Series 12	Internal data logging application	5 s
	Garmin Forerunner 970	BLE	0.25 - 0.5 s
	Google Pixel Watch 4	BLE	2 s
	Huawei Watch 5	BLE	1 s
	Oura Ring 5	HealthKit Export	5 s
	Samsung Galaxy Watch 8	Exported from companion app	1 s
	WHOOP 5	BLE	1 s
Background Mode	Apple Watch Series 12	Internal data logging application	5 s
	Garmin Forerunner 970	BLE	0.5 s
	Google Pixel Watch 4	BLE	2 s
	Huawei Watch 5	BLE	1 s
	Oura Ring 5	HealthKit Export	Variable (about every 300 s)
	Samsung Galaxy Watch 8	Exported from companion app	60 s (1 s during walking)
	WHOOP 5	BLE	1 s
Reference	Polar H10 ECG Chest Strap	BLE	1 s

Table 1. Heart Rate Reporting Interval and Collection Method by Device

Sample size was specified in advance as a minimum of 50 participants per workout activity per comparator device. Sizing, which pertains only to workouts and not Daily Living, assumed 80% power to detect a standardized paired difference (Cohen's d_z) of approximately 0.5 in Mean Absolute Error (MAE), using a two-sided test with $\alpha = 0.05$ and Bonferroni adjustment for the six device comparisons. The smallest detectable difference in mean MAE ranged from approximately 2.5 bpm (outdoor running and cycling) to 3.7 bpm (HIIT), under assumed between-participant MAE standard deviations of 5.0–7.5 bpm and a between-device correlation of 0.5. These assumptions were used for sizing only and the final analyses are described below.

Methods

Data quality checks were applied programmatically to the reference chest strap and test devices. For reference devices, this required adequate coverage (reading for ≥90% of the full session and ≥70% within an activity) and validity (no more than 5% of measurements fell outside 30–220 bpm). For test devices, this required a presence of a heart rate stream and the absence of extreme values (outside 20–240 bpm, a broad range because manufacturers' internal reporting bounds are not documented). Individual activities failing any checks were excluded from the affected comparison. Separately, individual reference samples outside 30–240 bpm were excluded from the analysis.

Matching the appropriate start time for each data session with that of the reference chest strap was completed using available timestamp data. To ensure devices were appropriately time synchronized using network time protocol (NTP), every device (including the Apple Watch and the Polar H10 chest strap) was Bluetooth®-paired to its own Wi-Fi-enabled companion phone. Study moderators confirmed alignment procedures were completed for all devices prior to the start of data collection and again at the time of data offload. Once aligned, each device's measurements were matched to the nearest-neighbor timestamp in the chest strap reference enabling a per-sample level accuracy evaluation against reference for each user experience designed by the manufacturer.

Data from Samsung Galaxy Watch 8 during daily activities (background) required different handling, as the export did not provide corresponding timestamping, only providing start and end timestamps, minimum and maximum heart rate, and a single additional heart rate value whose derivation is not documented. Testing this value against the reference at interval start, at interval end, and versus the mean of the reference across the interval, showed the closest agreement with the interval mean. This heart rate captured during daily activities was therefore compared against the mean of the chest strap reference over each interval.

Accuracy for each session was evaluated using both Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) calculations, which are standard methods of comparative accuracy testing, with RMSE being more sensitive in identifying a data stream that may have infrequent but significant errors. Each session-level metric (RMSE and MAE) for Apple Watch was compared directly with that of the comparative device. Beyond sample-level analysis, we also compared session-level average heart rate reported by each device against the reference chest strap, and then compared session-level accuracy between devices in each head-to-head evaluation. This allowed us to assess the accuracy of a session summary metric, which is especially relevant for devices without screens — where live heart rate cannot be provided to the user without engagement with a smartphone.

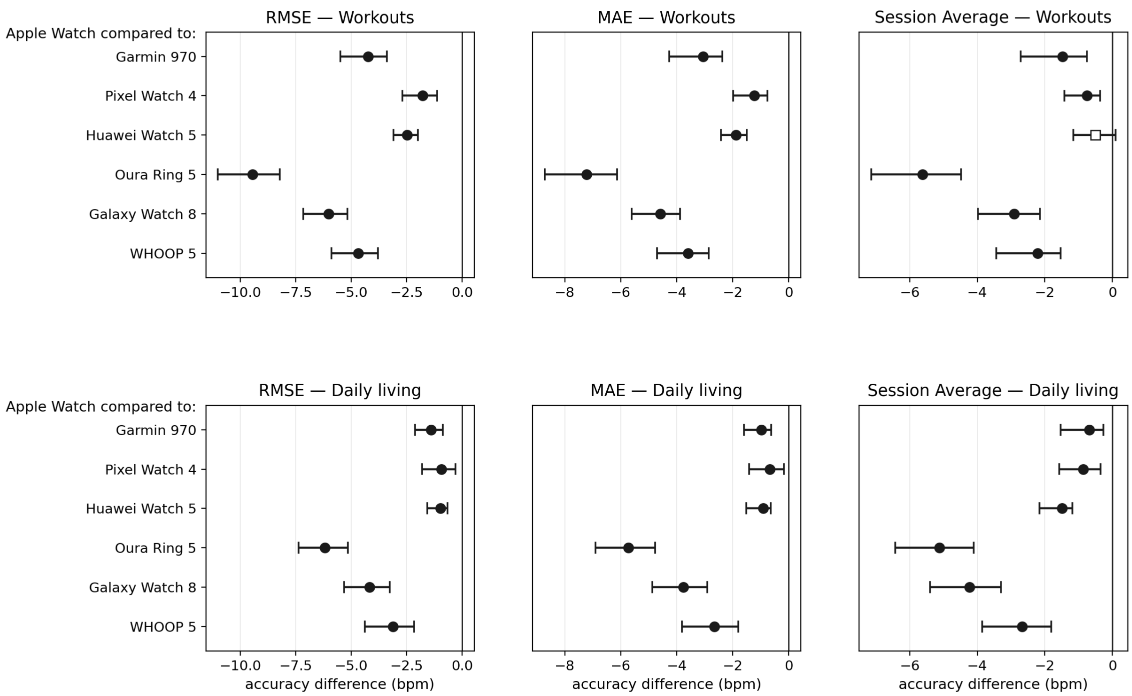
A linear mixed model (LMM) was used to model the paired difference (Apple Watch - Comparison Device) for each accuracy metric per participant and activity, treating activity as a fixed effect and incorporating a random intercept per participant to account for within-participant correlation of their repeated measures. The reported overall accuracy difference was the equally-weighted mean of the activity means, and a 99% bias-corrected and accelerated (BCa) bootstrap confidence interval (clustered by participant) was computed to account for the multiplicity of device comparisons. For mean error, the absolute error is taken per session before averaging, so the difference is a difference of mean absolute errors. Each reported metric is a mean accuracy difference (reported in bpm), so a negative difference means the error associated with Apple Watch was less than that of the comparison device.

Results

In total, the study enrolled 1460 participants, with representation across skin tone, sex, BMI, and age. Of these, 1254 contributed at least one paired Apple-versus-comparator measurement to the final analysis. Among those, 59% were enrolled at the third-party CRO-managed sites, 20% had a Fitzpatrick skin tone of V or VI, 35% were female, mean age was 39 ± 12 years, and mean BMI was 26 ± 5 kg/m².

Apple Watch demonstrated a statistically significant higher level of accuracy versus all other devices for both Overall Workout and Overall Daily Living (background) protocol for RMSE and MAE. Apple Watch also demonstrated better accuracy for session-level average heart rate for both Overall Workout and Overall Daily Living, in 11 of 12 comparisons, with the exception of Huawei Watch 5 for Overall Workout, where Apple Watch showed better numeric accuracy that was not statistically significant.

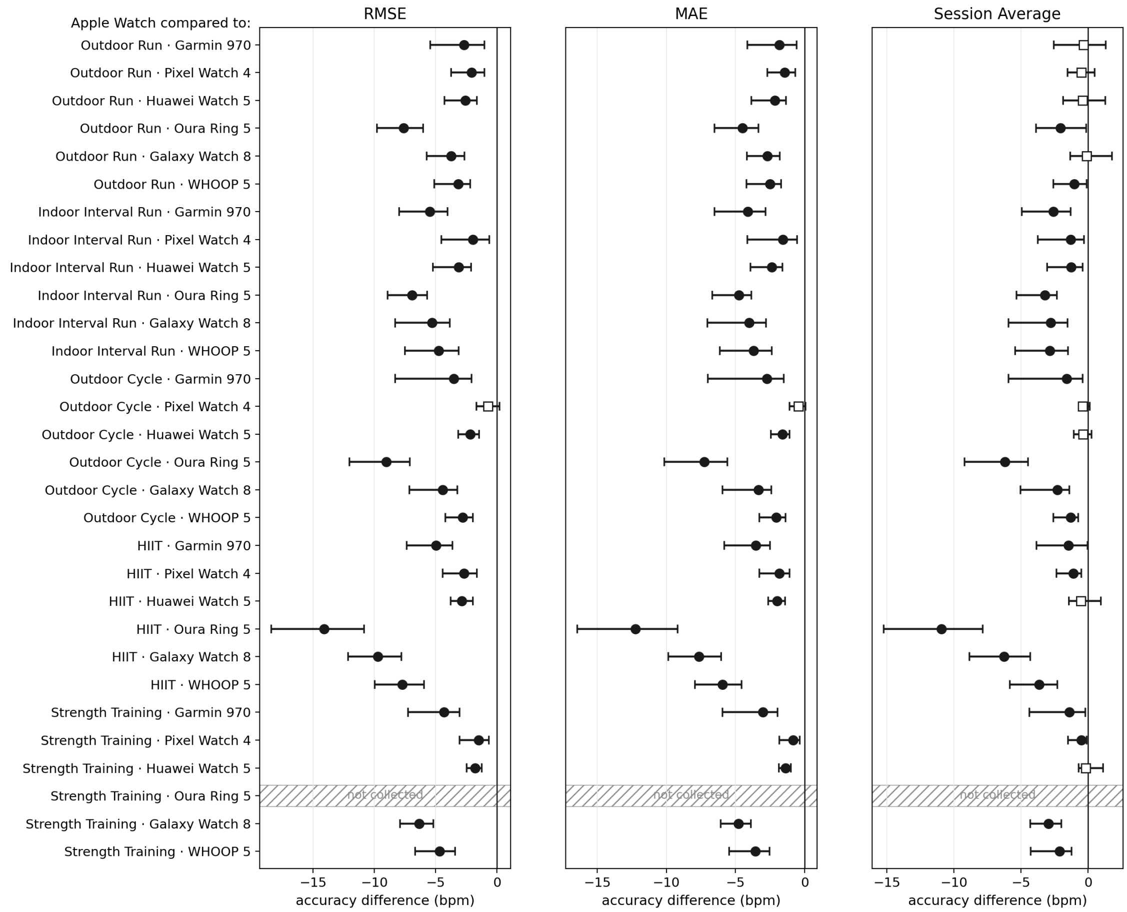
Results are reported below in Figures 1 through 3 for overall results, individual workout results, and individual daily activities results.



Accuracy difference with 99% confidence interval. A filled circle indicates a statistical difference and an open square indicates an indeterminate conclusion.

Figure 1: Overall Results (Apple Watch error minus comparison device error)

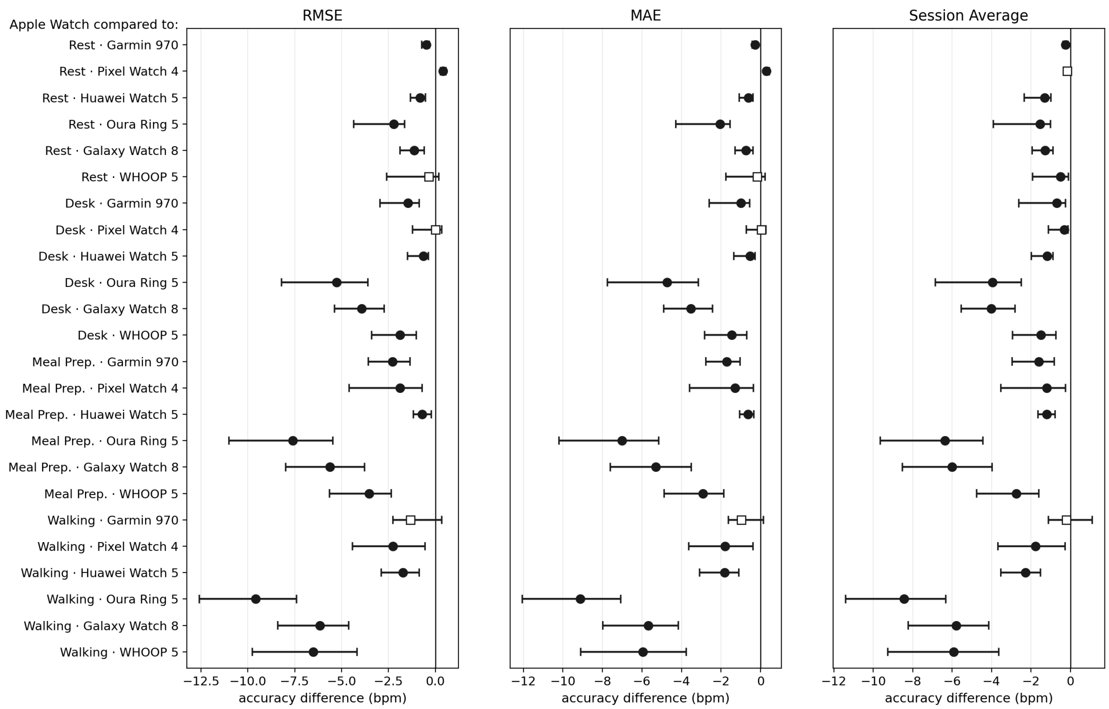
Apple Watch also showed statistically better accuracy for all individual workout activities (cycling, outdoor run, indoor interval run, HIIT, strength training) for all devices in RMSE and MAE, except versus Pixel Watch 4 in cycling, where the conclusions were indeterminate. For session level averages, Apple Watch showed better accuracy for the majority of devices in most activities, however in eight of the comparisons the conclusion was indeterminate.



Accuracy difference with 99% confidence interval. A filled circle indicates a statistical difference and an open square indicates an indeterminate conclusion.

Figure 2: Individual Workout Results (Apple Watch error minus comparison device error)

For individual activities within the Daily Living protocol, Apple Watch showed better accuracy for nearly all devices in all activities, across all metrics, with narrow exceptions. Eight conclusions were statistically indeterminate for rest, desk work, and walk. Pixel Watch 4 showed higher accuracy for RMSE and MAE for the rest condition only, however by less than a margin of 0.5 bpm.



Accuracy difference with 99% confidence interval. A filled circle indicates a statistical difference and an open square indicates an indeterminate conclusion.

Figure 3: Individual Daily Living Results (Apple Watch error minus comparison device error)

Across all individual activities spanning both the Workout and Daily Living protocols, there were 159 activity-by-measure comparisons. Apple Watch showed better accuracy with the 99% confidence level in 139, conclusions were indeterminate in 18, and the comparison device's accuracy was higher in 2. Evaluating the point estimate regardless of statistical significance, the comparison favored Apple Watch in 155 of 159.

Conclusion

Apple Watch demonstrated better heart rate accuracy versus all comparative devices in a large, statistically powered study, with broad evaluation of activities commonly practiced by smartwatch, smart ring, and fitness band users. Apple Watch offers users the most accurate heart rate sensing among the wearable devices evaluated in this study.

References

1. Chung V, Chopin L, Karadayi J, Grèzes J. 2026. "Validity of the Polar H10 for Continuous Measures of Heart Rate and Heart Rate Synchrony Analysis." *Sensors (Basel)*. Jan: 26(3):855. pmc.ncbi.nlm.nih.gov/articles/PMC12899868/.

2. Schaffarczyk M, Rogers B, Reer R, Gronwald T. 2022. "Validity of the Polar H10 Sensor for Heart Rate Variability Analysis during Resting State and Incremental Exercise in Recreational Men and Women." *Sensors (Basel)*. Aug: 22(17):6536. pmc.ncbi.nlm.nih.gov/articles/PMC9459793/.

3. Oura. "How to Choose the Right Oura Ring Size." Oura Support. Accessed August 2026. support.ouraring.com/hc/en-us/articles/360025590653-How-to-Choose-the-Right-Oura-Ring-Size.

4. Omdia. "Wearable Band Sell-In Shipments, 2Q26."

5. Oura. "General FAQs." Oura Support. Accessed August 2026. support.ouraring.com/hc/en-us/articles/4408961184147-General-FAQs.