

Supplementary Table 1. Previous studies comparing different smartwatches to assess sleep

Author (year)	Compared Smartwatches	Participants	Assessing period	Results
Montgomery-Downs et al. (2012) ⁷	Fitbit, PSG, Actiwatch	24 healthy adults	Laboratory, overnight	<p>Fitbit vs PSG: overestimate TST (67.1 min) overestimate SE (14.5%) 97.8% sensitivity (91.4% for N1, 97.7 for N2, 98.3 for N3, 98.8 for REM, 97.5% for arousal epochs) 19.8% specificity (42.9% for wake before sleep onset, 11.8% for WASO)</p> <p>Actiwatch vs. PSG: overestimate TST (43.0 min) overestimate SE (9.3%) 95.7% sensitivity (86.5% for N1, 96.1 for N2, 96.8 for N3, 95.9 for REM, 86.1% for arousal epochs) 38.9% specificity (35.8% for wake before sleep onset, 37.5% for WASO)</p> <p>Fitbit vs. Actiwatch significantly different TST: Fitbit overestimate TST (24.1 min) significantly different SE: Fitbit overestimate SE (5.2%)</p> <p>Fitbit One, Fitbit Zip, and Withings Pulse being the strongest performers.</p>
Ferguson et al. (2015) ²²	Seven consumer-level activity monitors (Fitbit One, Fitbit Zip, Jawbone UP, Misfit Shine, Nike Fuelband, Striv Smart Pedometer and Withings Pulse), and two research-grade accelerometers/multi-sensor devices (BodyMedia SenseWear and ActiGraph GT3X+)	21 healthy adults	48 hours in free-living condition.	<p>Consumer-level vs. S.research-grade: correlated strongly for TST</p> <p>Fitbit One, Fitbit Zip, and Withings Pulse being the strongest performers.</p>
Meltzer et al. (2015) ²³	Fitbit Ultra, PSG, 2 actigraphs (Ambulatory monitoring Inc. Motionlogger sleep watch, Phillips-Respironics Mini-Mitter Spectrum)	63 youth aged 3–27 years	Single night, sleep laboratory	<p>Fitbit vs. PSG: (normal mode) overestimate TST (41 min) underestimate WASO (32 min) overestimate SE (8%) 87% sensitivity 52% specificity 84% accuracy (sensitive mode) underestimate TST (105 min) overestimate WASO (106 min) underestimate SE (21%) 70% sensitivity 79% specificity 71% accuracy</p> <p>Fitbit vs. Ambulatory monitoring Inc. (normal mode) overestimate TST (37 min) underestimate WASO (32 min) overestimate SE (7%). (sensitive mode) underestimate TST (98 min) overestimate WASO (101 min) underestimate SE (19%).</p> <p>Fitbit vs. Phillips-Respironics Mini Mitter actiwatch spectrum: (normal mode) overestimate TST (36 min) underestimate WASO (32 min) overestimate SE (7%). (sensitive mode) underestimate TST (117 min) overestimate WASO (109 min) underestimate SE (23%).</p> <p>For TST: No difference among all devices; Strong correlation with PSG.</p> <p>For SE: Withings, Misfit, Fitbit, Basis differed from PSG while Actiwatch did not.</p> <p>Light sleep: All devices underestimate light sleep compared to PSG.</p> <p>Deep sleep time: Basis did not differ from PSG while Misfit and Withings differed.</p>
Mantua et al. (2016) ¹³	PSG, five wearable devices (Basis Health Tracker, Misfit Shine, Fitbit Flex, Withings Pulse O2, and a research-based actigraph, Actiwatch Spectrum)	40 healthy adults aged 18–30 years	Home-based, n.r. (overnight?)	<p>For TST: No difference among all devices; Strong correlation with PSG.</p> <p>For SE: Withings, Misfit, Fitbit, Basis differed from PSG while Actiwatch did not.</p> <p>Light sleep: All devices underestimate light sleep compared to PSG.</p> <p>Deep sleep time: Basis did not differ from PSG while Misfit and Withings differed.</p>
Dickinson et al. (2016) ²⁴	Fitbit Charge HR, Actiwatch	38 young adults	4 nights	<p>Fitbit vs. Actiwatch: overestimate TST 88% sensitivity for poor sleepers, 98% sensitivity for good sleepers</p>
Osterbauer et al. (2016) ²⁵	Fitbit, PSG	14 children aged 3–11 years with sleep disordered breathing	Single night	<p>Fitbit vs. PSG similar TST (r=0.99) underestimate TST (20.5 min) overestimate SE (8%) 99% sensitivity 10% specificity</p>
Rosenberger et al. (2015) ²⁶	Fitbit One, Jawbone Up, GENEactive, GT3X+, and Z-machine.	40 adults	Free-living, 24 hours	<p>Fitbit, Jawbone, GENEactive, GT3X+ vs. Z-machine: GT3X+ overestimate TST (4 min) Jawbone Up overestimate TST (32 min) Fitbit overestimate TST (36 min) GENEactive underestimate TST (36 min)</p>
Brooke et al. (2017) ¹⁴	8 monitors: Nikes+ FuelBand SE, Garmin VivoFit, Misfit Shine, Fitbit Flex, Jawbone UP, Polar Loop, Fitbit Charge HR, and SenseWear Armband Mini. (Criterion: a sleep log)	95 healthy adults	Free-living, One night	<p>Mean absolute percent errors for TST: Garmin VivoFit (4.0%) Fitbit Flex (8.8%), Jawbone UP (10.2%) Fitbit Charge HR (11.5%) Misfit Shine (12.9%) SenseWear Armband Mini laying down (13.6%) Polar Loop (17.5%) SenseWear Armband Mini (21.61%)</p>
Cook et al. (2017) ²⁰	Fitbit Flex, PSG, Actiwatch	21 unipolar major depressive disorder patients	Laboratory condition, overnight	<p>Fitbit vs. PSG: (normal mode) overestimate TST (46 min) underestimate WASO (44 min) overestimate SE (8.1%) similar SOL (2 min) 98% sensitivity 35% specificity 88% accuracy</p> <p>(sensitive mode) underestimate TST (86.3 min) overestimate WASO (74.8 min) underestimate SE (16%) overestimate SOL (11.5 min) 78% sensitivity 80% specificity 78% accuracy</p> <p>Actiwatch vs. PSG: overestimate TST (40.6 min) underestimate WASO (27.1 min) underestimate SE (7%) underestimate SOL (13.5 min) 97% sensitivity 31% specificity 87% accuracy</p>
Lee et al. (2017) ²⁰	Fitbit charge HR, Actiwatch 2 (sleep log was used to edit Actiwatch data)	16 healthy young adults	14 days, no data for each sleep phase	<p>Fitbit vs. Actiwatch: overestimate TST (22.4 min) not different sleep onset time</p>
Castner et al. (2019) ²⁸	Fitbit Charge TM, Actigraph wGT3X+	47 women with asthma	2 weeks	<p>Fitbit vs. Actigraph: underestimate TST underestimate wake counts overestimate SE 97% sensitivity 40% specificity</p>
de Zambotti et al. (2018) ¹⁵	Fitbit charger 2, PSG	44 adults without mental or medical condition	One overnight	<p>Fitbit vs. PSG: overestimate TST (9 min) and light sleep (34 min) underestimate deep sleep (24 min) underestimate SOL (4 min) not significantly different WASO and REM 96% sensitivity 61% specificity, 81% accuracy for light sleep, 49% for deep sleep, 74% for REM</p>
Feehan et al. (2018) ¹⁹	Fitbit devices	Systematic review	Not applicable	<p>In controlled settings: (normal-mode) overestimate TST and SE by more than 10%. (sensitive mode) underestimate TST and SE by more than 15%.</p> <p>In free-living settings: overestimate TST compared with a portable sleep monitor (MAPE approximately 10%) or Actiwatch accelerometer (approximately 10 minutes per night)</p>
Hakim et al. (2022) ²⁹	Fitbit Charge PSG	22 children aged 3–18 years	Laboratory, Single night	<p>Fitbit vs. PSG: overestimate TST (30 min). underestimate total wake time (23 min)</p>
Liang et al. (2018) ²⁰	Fitbit Charge 2, Wearable EEG-based eye mask Neuroon, Medical sleep monitor (Sleep Scope)	25 healthy adults	Free-living condition, 3 nights	<p>Fitbit vs. medical sleep monitor: underestimate TST, light sleep, REM overestimate deep sleep no significantly different SE overestimate WASO underestimate SOL</p> <p>Neuron vs. medical sleep monitor: underestimate TST, light sleep overestimate deep sleep no significantly different REM underestimate SE overestimate WASO overestimate SOL</p>
Weatherall et al. (2018) ²⁰	Fitbit Charge HR, patient-reported outcomes for sleep patterns (3 times during 2 weeks)	86 adults with type 2 diabetes mellitus	14 days	<p>Fitbit vs. patient report: Weakly correlated for troubled sleep nights and minutes asleep</p>
Haghighyeh et al. (2019) ²⁷	Wristband Fitbit	Systematic review	Not applicable	<p>Fitbit: overestimate TST (7–67 min) overestimate SE (2–15%) underestimate WASO (6–44 min) No significant different SOL 81%–91% accuracy 87%–99% sensitivity 10%–52% specificity.</p> <p>Fitbit vs. PSG: underestimate TST (24–47 min) overestimate WASO (21–41 min) ≥90% sensitivity (≥68% for N1+N2, ≥50% for N3, ≥72% for REM) ≥88% specificity Inconsistent underestimate or overestimate for N1+N2 underestimate N3 sleep (21–46 min)</p> <p>Actiwatch2 vs. PSG: (medium setting) underestimate TST (24–38 min) overestimate WASO (22–38 min) 90%–91% sensitivity 80%–86% specificity 89%–90% accuracy (high motion threshold) underestimate TST (7–12 min) overestimate WASO (11–18 min) 95% sensitivity 64%–73% specificity 93%–94% accuracy</p>
Lee et al. (2019) ³⁰	Fitbit Alta HR, Actiwatch 2, PSG	58 healthy adolescents aged 15–19 years	5 nights	<p>Fitbit vs. PSG: underestimate TST (24–47 min) overestimate WASO (21–41 min) ≥90% sensitivity (≥68% for N1+N2, ≥50% for N3, ≥72% for REM) ≥88% specificity Inconsistent underestimate or overestimate for N1+N2 underestimate N3 sleep (21–46 min)</p> <p>Actiwatch2 vs. PSG: (medium setting) underestimate TST (24–38 min) overestimate WASO (22–38 min) 90%–91% sensitivity 80%–86% specificity 89%–90% accuracy (high motion threshold) underestimate TST (7–12 min) overestimate WASO (11–18 min) 95% sensitivity 64%–73% specificity 93%–94% accuracy</p>
Liang et al. (2019) ³¹	Fitbit charge 2, medical device (Sleep Scope)	23 adults without chronic conditions	Free-living condition, 3 nights	<p>Fitbit vs. medical device: underestimate probabilities for staying in a sleep stage underestimate for transitioning to another stage</p>
Liu et al. (2019) ³⁶	Fitbit Alta, Consensus sleep diary	10 healthy adults aged 20–24 years	Free-living condition, 7 consecutive days	<p>Fitbit vs. diary: underestimate TST (5.46 min) underestimate SE (1.35%) overestimate SOL (4.99 min) overestimate WASO (13.09 min)</p>
Moreno-Pino et al. (2019) ⁴¹	Fitbit (Fitbit charge 2 or Fitbit Alta HR), PSG	65 adults (55 with obstructive sleep apnea)	Laboratory condition, overnight	<p>Fitbit vs. PSG: overestimate TST (59.8 min, significant), underestimate WASO (36.1 min, significant), underestimate SOL (23.2 min, significant) overestimate light sleep (68.8 min, significant) underestimate deep sleep (74.2 min, significant) overestimate REM (2.8 min, not significant) underestimate wake (41.9 min, significant) 87.8% sensitivity 43.9% specificity</p> <p>In normal PSG: 89.3% sensitivity, 39.5% specificity</p> <p>In OSA, 87.6% sensitivity, 44.8% specificity</p> <p>In OSA under CPAP treatment: 88.2% sensitivity, 55.1% specificity</p> <p>Fitbit vs. Actigraph: overestimate TST (5.7 min), moderate correlation (r=0.67) 10.1% mean absolute percentage error overestimate WASO (0.21 min), poor correlation (r=0.32)</p> <p>Garmin vs. Actigraph: overestimate TST (55.39 min), poor correlation (r=0.42) underestimate WASO (35.86), no correlation (r=0.01)</p>
Haghighyeh et al. (2020) ²⁴	Fitbit charge 2TM, actigraphy (Motionlogger*, Micro Watch Actigraph; MMWA), electroencephalography (EEG; Zmachine)	35 healthy adults	Free-living condition, single night	<p>Fitbit vs. EEG: overestimate TST (17 min, not significant) underestimate WASO (6.5 min, not significant) underestimate SOL (10.5 min, significant) overestimate SE (3.8%, significant) underestimate REM (7.5 min, not significant) overestimate light sleep (52.9 min, significant) 54% wake/sleep agreement 95% sensitivity 97% specificity</p> <p>Actigraph (MMWA) vs. EEG: overestimate TST (16.1 min, not significant) underestimate WASO (6.4 min, not significant) underestimate SOL (9.7 min, significant) overestimate SE (3.7%, not significant) 51% wake/sleep agreement 94% sensitivity 53% specificity</p> <p>Fitbit vs. Actigraph (MMWA): No significant difference in SOL, WASO, TST, SE Better minimum detectable change for Fitbit than actigraphy (MMWA)</p>
Godino et al. (2020) ²²	Fitbit Charge HR, PSG	26 healthy children aged 9–11 years	Overnight	<p>Fitbit vs. PSG: underestimate TST (14 min) overestimate WASO (9 min). 95.7% sensitivity 56.9% specificity 92.1% accuracy</p>
Thota (2020) ³⁵	Fitbit Flex, daily self-reported mood	17 healthy adults	2 weeks	<p>Fitbit vs. diary: moderate correlation (r=0.643, significant)</p>
Burkart et al. (2021) ³²	Fitbit Charge 3 Actigraph GT19X PSG	56 children (79% had obstructive sleep apnea)	Single night, laboratory	<p>Fitbit vs. PSG: underestimate TST (6.1 min) overestimate EF (9%) overestimate WASO (18.8 min) underestimate sleep onset (1.2 min) overestimate sleep offset (6 min) underestimate WASO and SE at high values.</p> <p>ActiGraph vs. PSG: underestimate TST (31.5 min) underestimate SE (12.9%) overestimate WASO (56.1 min) underestimate sleep onset (10.2 min) overestimate sleep offset (10.5 min)</p>
Chinoy et al. (2021) ²³	PSG, Actigraphy (Philips Respironics Actiwatch 2), 4 wearable (Fatigue science readiband, Fitbit Alta HR, Garmin Fenix 5S, Garmin Vivomart 3), 3 nonwearable (EarlySense Live, ResMed S+, SleepScore Max)	34 healthy young adults	3 nights in sleep laboratory	<p>Actiwatch vs. PSG: overestimate TST (23.9 min, significant) overestimate SE (5.0 min, significant) underestimate SOL (7.6 min, significant)</p> <p>ActiGraph vs. PSG: overestimate TST (15.3 min) overestimate SE (2.8 min) underestimate SOL (0.7 min, significant)</p> <p>ActiGraphFitbit Alta HR vs. PSG: overestimate TST (2.6 min) overestimate SE (0.9 min) underestimate SOL (3.1 min, significant)</p> <p>ActiGraphGarmin vs. PSG: overestimate TST (43.7–46.8 min, significant) overestimate SE (10.1–10.6 min, significant) overestimate SOL (0.8 min, significant) or underestimate (Vivomart3, 1.1 min)</p> <p>ActiGraphEarlySense vs. PSG: overestimate TST (13.6 min, significant) overestimate SE (2.9 min, significant) overestimate SOL (0.8 min, significant)</p> <p>ActiGraphResMed vs. PSG: underestimate TST (0.3 min) similar SE overestimate SOL (4.0 min, significant)</p> <p>ActiGraphSleepScore vs. PSG: overestimate TST (7.5 min) overestimate SE (1.6 min) overestimate SOL (4.4 min, significant)</p>
Menghini et al. (2021) ³¹	Fitbit Charge 3, PSG	39 adolescents aged 16–19 years (12 with insomnia)	Single night	<p>Fitbit vs. PSG: underestimate TST (11 min) underestimate TST (2.5 min) overestimate WASO (9 min) underestimate light and deep sleep duration proportionally higher discrepancies of TST and SE for longer sleep hours Sensitivity 68% for wake, 78% for light sleep, 59% for deep sleep, 69% for REM.</p>
Stucky et al. (2021) ³⁵	Fitbit Charge 2 PSG	59 shift workers	One month for Fitbit Charge 2, two nights for PSG	<p>Fitbit vs. PSG: overestimate TST (4.0 min) overestimate REM duration (2.7 min) overestimate light sleep (10.4 min) underestimate deep sleep (11.2 min) overestimate REM latency (29.4 min, significant) overestimate WASO (37.1 min, significant) Sensitivity 42.8% for WASO, 53.4% for light sleep, 27.9% for deep sleep, 54.8% for REM sleep (43.2% for REM sleep shorter than 120 min, 57.0% for longer than 120 min) Specificity 89.8% for WASO, 57.4% for light sleep, 92.0% for deep sleep, 88.9% for REM sleep (96.3% for REM sleep shorter than 120 min, 86.4% for REM sleep longer than 120 min) Accuracy 82.4% for WASO, 55.3% for light sleep, 77.6% for deep sleep, 86.1% for REM sleep (93.4% for shorter, 83.7% for longer than 120 min)</p>
Dong et al. (2022) ³⁴	Fitbit Charge 4 PSG actigraphy	37 participants with chronic insomnia disorder	One night, sleep laboratory	<p>Fitbit vs. PSG: underestimate TST (11 min) underestimate SE (4.9%, significant) underestimate sleep onset latency (1.8 min) overestimate WASO (2.8 min) overestimate light sleep (37.7 min, significant) underestimate deep sleep (41.4 min, significant) underestimate REM (4.7 min) 86.5% accuracy 89.9% sensitivity 62.2% specificity</p> <p>Actiwatch vs. PSG: overestimate (2.3 min) underestimate SE (1.5%) underestimate sleep onset latency (1.0 min) underestimate WASO (14.8 min) 86.9% accuracy 92.6% sensitivity 35.7% specificity</p> <p>Actiwatch vs. Fitbit: Actiwatch overestimate TST (13.8 min, significant) Actiwatch underestimate SE (3.6%, significant) Actiwatch underestimate SOL (0.8 min) Actiwatch overestimate WASO (19.2 min, significant) 87.9% accuracy 89.1 sensitivity 75.7 specificity</p>
Kawasaki et al. (2022) ⁴⁴	Fitbit Alta HR, EEG	40 college athletes	2 nights, Laboratory	<p>Fitbit vs. EEG: strong correlation in TST (r=0.83) modest correlation in N3 (r=0.68) strong correlation in the percentage of N3 between sleep onset and initial REM (r=0.73)</p>
Lim et al. (2023) ³⁶	Fitbit inspire 2, PSG	9 participants without sleeping disorders		<p>Fitbit vs. PSG: overestimate TST, deep sleep, REM (significant) overestimate time in bed, sleep efficiency, wake after sleep onset (not significant) 93.9% sensitivity 13.1% specificity 76.0% accuracy</p>
Byun et al. (2023) ³⁰	Fitbit charge 2 Galaxy watch 2 PSG	127 obstructive sleep apnoea patients		<p>Fitbit vs. Galaxy: not significantly different recording failure rate</p> <p>Fitbit, Galaxy vs. PSG: Smartwatches underestimate TST (Fitbit: 27.5 min, Galaxy: 24.9 min)</p>

PSG, polysomnography; TST, total sleep time; SE, sleep efficiency; REM, rapid eye movement; SOL, sleep onset latency; WASO, wake after sleep onset; MAPE, mean or median absolute percentage error