



Activity Trackers Revisited

7th July 2026

We'll cover today...

1. Which metrics are useful
2. Which metrics are misleading or inaccurate
3. Which metrics can actually make you less healthy if you obsess over them
4. Which devices are best for which metric

Metric #1: Steps

A systematic review of 16 research studies found that **step counting is consistently one of the most accurate things wearables do.**

- **Garmin:** ~83% accuracy
- **Apple:** ~81% accuracy
- **Fitbit:** 77.3% accuracy
- **Polar:** 53% accuracy

Metric #1: Steps

Good uses:

- ✓ Create a minimum activity target
- ✓ Monitor overall activity
- ✓ Prevent activity decline when working-from-home
- ✓ Track trends over months

Bad uses:

- ✗ Obsessing getting over exactly 10,000 steps
- ✗ Walking around the kitchen at 9pm to hit a target

Takeaway: 9,200 steps and 10,000 steps are physiologically identical. Averaging 3,500 a day when you're aiming for 7,500 is a problem!

Metric #2: Heart Rate

Wrist-based heart rate monitoring has improved massively:

- Apple Watch consistently ranks highest (86.31 % accuracy)
- Fitbit next best (73.56%)
- Garmin performs ok (67.73% accuracy, mine also seems unreliable)
- Chest straps remain the gold standard (99% accuracy)

Accuracy falls during: Intervals, Strength training, Movements involving gripping/flexing the wrist i.e. strength training again!

Metric #2: Heart Rate

Good uses:

Resting HR - one of the best metrics available for judging whether you should push or pull back.

Can indicate: improving fitness (resting HR dropping steadily), accumulated fatigue, illness (sharp increase), stress, poor sleep, struggling with altitude and more.

Example:

"My resting heart rate is normally 39-40 but this week it's 41."

Small changes can be significant - get to know *your* baseline.



Metric #3: HRV

HRV = Heart Rate Variability (measured in milliseconds)

- Reflects how adaptable and recovered your nervous system is
- Higher isn't always better, but usually is!
- Massive variation from person to person (25 avg. to 300 avg.)
- HRV varies due to: alcohol, illness, menstrual cycle, menopause symptoms, poor sleep, psychological stress, travel, hard training and even eating patterns!



Metric #3: HRV

Look at HRV trends, not the specific number e.g.

Situation	HRV trend
Getting fitter	Gradual increase
Overstressed	Gradual decline
Illness/run down	Sudden drop
Drinking alcohol	Temporary drop
Better sleep	Improvement
Big altitude change	Sudden drop



Note: sometimes it spikes then drops showing something is off too!

Metric #4: Sleep

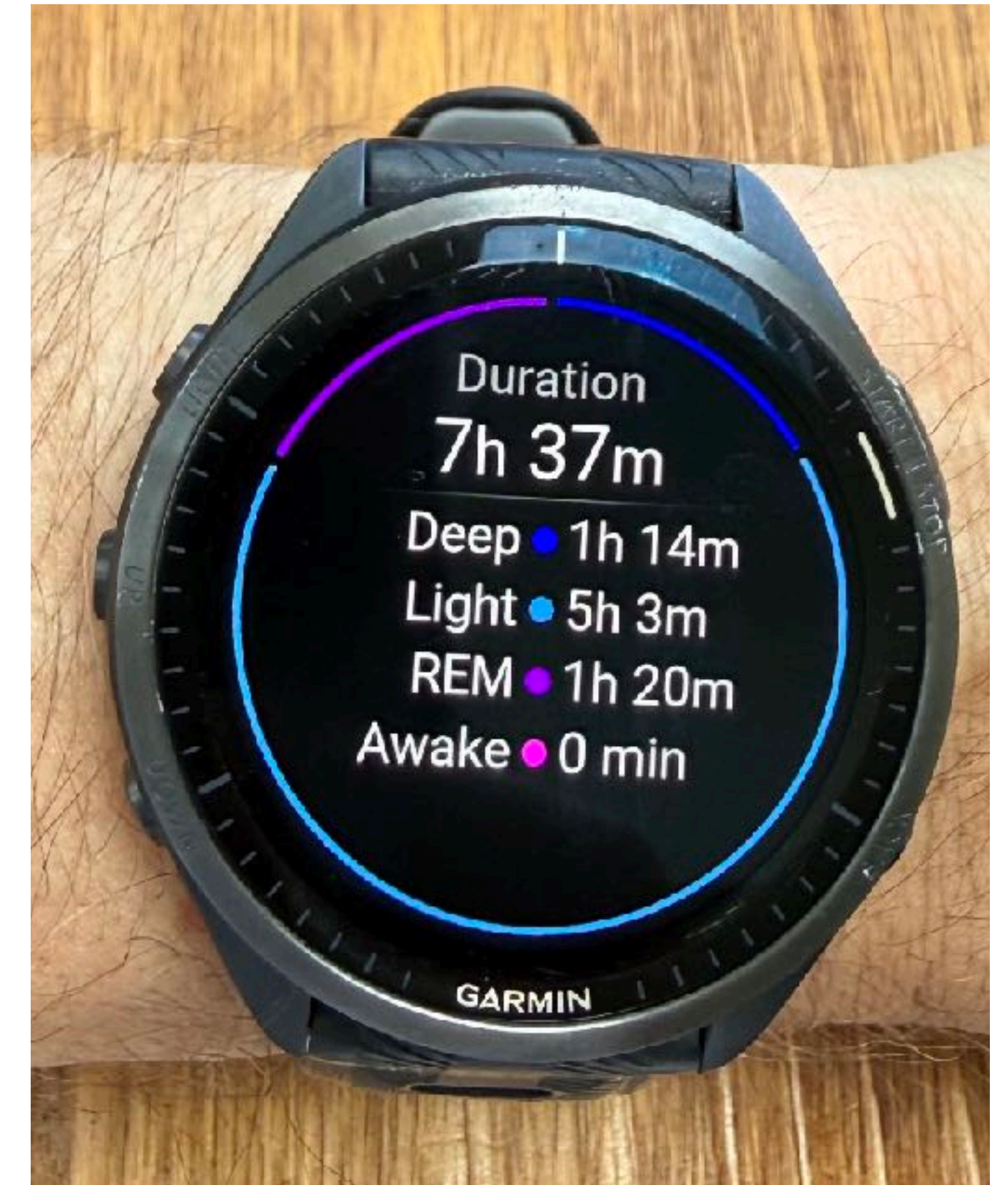
Sleep duration e.g. 7h 13 mins:

✓ Reasonably accurate.

Sleep stages (hours of light, REM, deep etc...):

✗ Not very accurate at all.

Large validation studies comparing wearables against sleep laboratories (polysomnography) found only fair-to-moderate agreement for sleep stage classification.



Metric #4: Sleep

Ask: “How rested do I actually feel?”

Plus factor in total sleep time, sleep consistency (how many days with enough hours of sleep) and how you felt you slept.

Example:

Client A:

“Poor” sleep score but feels
fantastic & has energy

Result: Ignore the watch

Client B:

Sleep score poor
Feels terrible, craving sugar,
exhausted & can't focus

Result: The watch is confirming what
you already knew.

Metric #5: Calories burned

The worst metric by far!

Errors of 20-50% are common e.g.

You burn: 350 kcal (if measured in a lab)

Watch says: 520 kcal

You “eat back” 500 calories and wonder why you didn’t lose any weight (or even gained it over time!).



Metric #5: Calories burned

Never eat because your watch says you earned it.

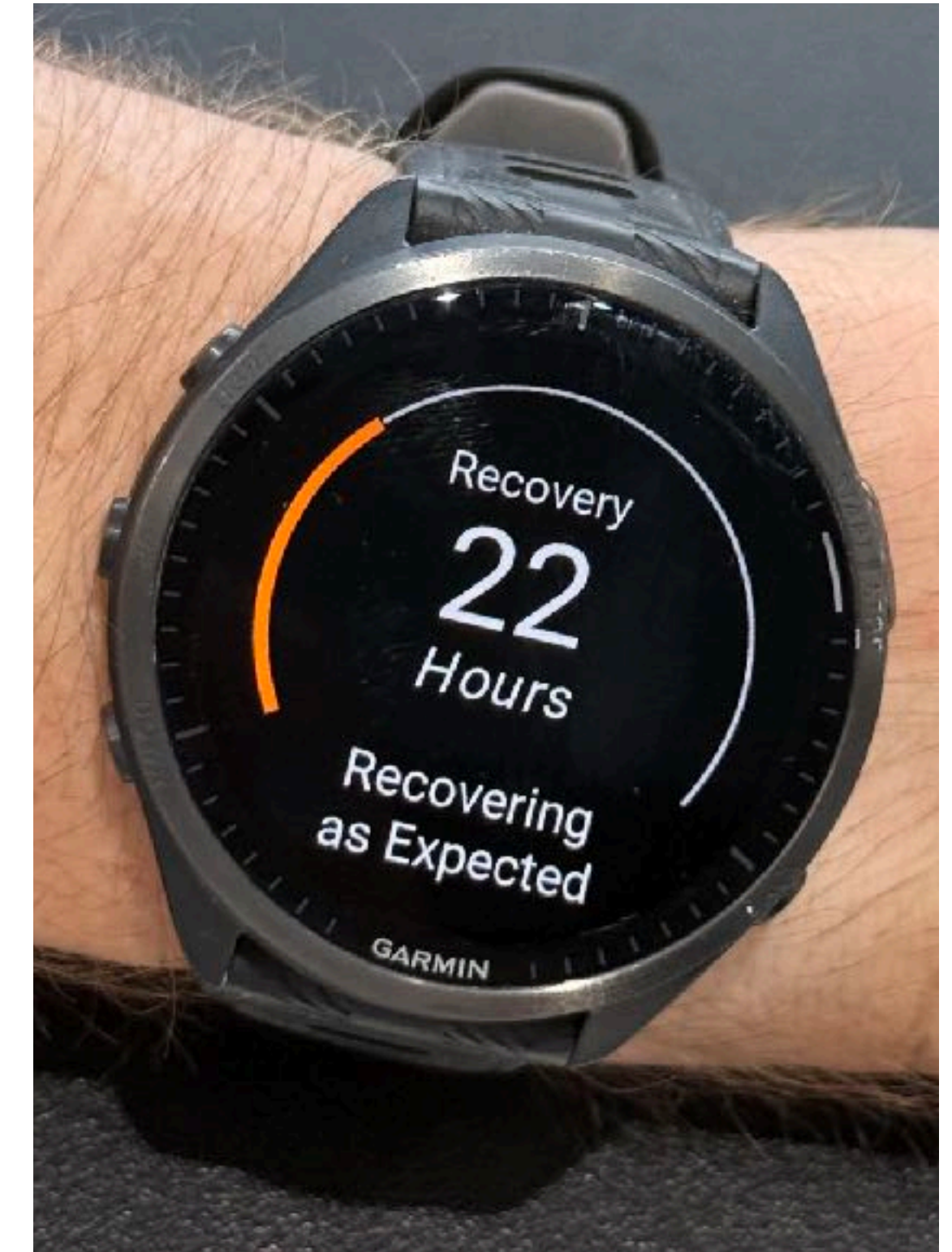
Instead:

- follow nutrition targets
- use bodyweight trends
- use hunger/fullness signals (some hunger is normal when losing weight!)
- use performance (e.g. lifting heavier, running faster etc...)

Metric #6: Readiness Scores, Recovery Scores & Stress Scores

Examples:

- Oura Readiness
- WHOOP Recovery
- Garmin Body Battery
- Fitbit Readiness



Metric #6: Readiness Scores, Recovery Scores & Stress Scores

These are educated guesses created from multiple imperfect measurements.

They can be useful for spotting trends but should never override:

- how you feel
- common sense
- your training plan

In summary...

TRINITY
TRANSFORMATION

Wearables are generally very good at measuring trends and behaviours, reasonably good at measuring some physiological variables, and surprisingly poor at estimating things like calories burned and sleep stages.

The wearable hierarchy

 **TRUST A LOT**

 **USEFUL WITH CAUTION**

 **DON'T RELY ON TOO MUCH**

Steps

HRV (Heart Rate Variability)

Calories burned

Resting heart rate

Sleep duration

Sleep stages (deep sleep, REM etc.)

Exercise heart rate

Readiness/recovery scores

Single-day HRV changes

Long-term trends (weeks/months)

Stress scores/body battery

Exact recovery percentages

Changes from your own baseline

Sleep scores

"You are recovered/unrecovered" notifications

TRINITY
TRANSFORMATION

Bonus: Device Summary

Biometric	 Winner	 Second	 Third	Worst
Sleep Staging	Oura ($\kappa=0.65$)	Apple ($\kappa=0.60$)	Fitbit ($\kappa=0.55$)	—
Deep Sleep	Oura (79.5%)	Fitbit (61.7%)	Apple (50.5%)	—
Wake Detection	Oura (68.6%)	Fitbit (67.7%)	Apple (52.4%)	Garmin (27%)
Nocturnal HRV	Oura (MAPE 5.96%)	WHOOP (8.17%)	Garmin (10.52%)	Polar (16.32%)
Active HR	Apple (86.3%)	Fitbit (73.6%)	Garmin (67.7%)	—
Step Count	Garmin (82.6%)	Apple (81.1%)	Fitbit (77.3%)	Oura (poor)
SpO2	Apple (MAE 2.2%)	Garmin Fenix (~4.5%)	Withings (~4.8%)	Garmin Venu (5.8%)
Calories	Apple (71%)	Fitbit (65.6%)	—	Garmin (48%)

K = Cohen's Kappa, a statistical metric used to measure inter-rater or intra-rater reliability for categorical data. It quantifies the level of agreement between two raters classifying the same items while removing the proportion of agreement that could occur purely by random chance