IJEMD-SS, 5 (6) (2026) https://doi.org/10.54938/ijemd-ss.2026.05.6.887

International Journal of Emerging Multidisciplinaries: Social Science

Research Paper
Journal Homepage: www.ojs.ijemd.com
ISSN (Online): 2958-0277



Energy-Matched Heart-Rate Sensing Schedules for Pre-Symptomatic Infection Detection on Low-Cost Wearables

1. Hassan Jalal, 2. Muhammad Usman Riaz

1. A-Level Student and Student Researcher, International School Lahore, Lahore, Pakistan

hassanjalalahmad2@gmail.com

ORCID: https://orcid.org/0009-0006-0943-7139

2. BSPT (UHS), MS-OMPT (RIU), PhD scholar in Rehabilitation Sciences, Superior University, Lahore, Pakistan

Associate professor/ HOD Physiotherapy Department FMH/ Principal FMH-IAHS

Advisor to Group Head Operation NUR-FMS

m.usman_711@live.com

ORCID: https://orcid.org/0000-0002-5287-9665

Abstract

Most wearable infection-detection studies have assumed that more sensing is better, and focused on deviations in resting heart rate preceding self-reported symptoms; however, not much consideration has been given to when sensing is needed when there is a limited budget for the amount of energy the device can expend. This study asks: Is there a difference in pre-symptomatic warning performance when the timing of the heart-rate acquisition is considered in addition to the amount of sensing? Two publicly released data sets (Stanford Phase 1 and Stanford Phase 2) were re-analyzed using a preregistered causal pipeline. The final frozen base cohort included 38 participants; 30 contributed to the primary paired comparison after arm-specific calibration requirements. On the primary, modelled-energy budget (E3), sparse schedules were employed with 7 acquisition bursts per day of 10 minutes each. The confirmatory comparison tested nocturnally sensing (S3) against uniform block (S2) through

the personalized time-of-day-normalized resting-heart-rate detector, with an identical alert burden (no more than two alert-days/person-month) and with a hierarchical paired endpoint of detection and warning time. S3 did not outperform S2: W/L/T = 8/12/10, θ = 0.433 (95% BCa CI 0.300-0.583), exact two-sided p = 0.5034. Detection rate was 63.3% for S3 and 66.7% for S2 and median warning time was 13 and 14 days respectively for the detected cases. Due to H1 failure, the fixed-sequence gate was closed and the remaining schedule comparisons were estimation-only. For all sensitivity analyses reported in H1, θ was kept at or below 0.5. These results do not demonstrate equality, but they do not offer any evidence that heart-rate acquisition focused at night is superior to an energy-matched uniform heart-rate acquisition schedule in warning about infection prior to symptoms onset in this sample. The study presents a framework for evaluation of sensing placement, alert burden, missingness and modeled device energy that is reproducible and suggests the need for larger prospective validation and measured hardware-energy experiments.

Keywords: wearable sensing, heart rate, infection detection, pre-symptomatic warning, energy-aware sensing, duty cycling, resting heart rate, smartwatch

1. Introduction

The early detection of infectious disease is useful as the physiological changes may occur prior to disease recognition and/or seeking of testing. Consumer wearables are appealing here because they passively capture heart rate and activity data over extended periods of time, and can be used to create a baseline for each individual to which deviations can be compared. Earlier, population-level studies indicated that wearable resting heart rate (RHR) and activity data is effective for tracking influenza-like illness (ILI), and studies on COVID-19 have also shown that physiological changes associated with COVID-19 can be detected both retroactively and prospectively (Radin et al., 2020; Mishra et al., 2020; Quer et al., 2021; Alavi et al., 2022). In Mishra et al., physiological changes were detected in 26 out of 32 COVID-19 cases analyzed, of which 22 out of 25 cases with available symptom data had the alterations detected prior or at the onset of the symptoms. Using a real time-like resting heart-rate alert retrospectively identified 63% of cases before the appearance of symptoms. Later, Alavi et al. reported that they sent alerts at or before the onset of signs or diagnosis in 67 of 84 confirmed cases, with a median pre-symptomatic warning time of three days. While such results provide proof that wearable physiology can include valuable early warning information, they on their own do not define an ideal acquisition schedule for a resource-constrained wearable.

The acquisition schedule is important since optical and activity sensing, computation and radio communication are consuming energy. Unfortunately, continuous sampling is not always possible with a low-cost device without decreasing battery life, increasing the size of the device, or increasing the frequency of charging. This results in an observation-control problem: a fixed energy budget can either be spread evenly throughout the day or focused into specific time periods or used randomly, or used at a point of estimated rest. Quickest-change-detection theory has long known that observation cost and detection delay should not be looked at in isolation, and every observation cannot simply be assumed to be free (Banerjee and Veeravalli, 2012). In wearable systems the same principle is not abstract, but actually practical: if it is possible to only use a limited number of acquisition bursts, they should be placed in such a way that early-warning performance is maintained.

A night time sensing is a reasonable suspect. During sleep and inactivity, resting heart rate is lower, and is less affected by movement artifacts, and there are significant differences in resting heart rate between day and night in wearable studies. For instance, Speed et al. (2023) found a decreased resting heart rate during the night and the lowest resting HR were typically reported between 03:00 and 07:00. However, sleep and circadian research warns that the time of day and measurement protocol may have significant consequences on wearables cardiac measurements. But a night time boost may be possible for two reasons at least. First, night could simply be rest, and contain cleaner observations of the heart-rate. Second, there may be further information about the time of day following gross rest-state matching. Explanation of these can easily be mixed up if it is compared with only an unrestricted daytime schedule.

The second methodological issue is that analyses of wearable infection-detection data are susceptible to leakage, confounding and missingness. Cleary et al. (2022) demonstrated how behaviour following a COVID-19 test result can itself impact on wearable signals, providing a pathway whereby post-test behaviour could be mistaken as detection of disease. Baseline construction may also lead to over-reporting of performance when information from one period of evaluation or from one stream at full resolution is leaked into sparse arms. More broadly, a recent study by Tognotti et al. 2026 showed that baseline-normalisation decisions can actually inflate the actual performance of the wearable classification when temporal boundaries are not respected. Then there's the missing data aspect: recent research has demonstrated that the performance of disease detection can be sensitive to missing data and the imputation method

used (Wallner et al., 2026). These concerns drive a causal, schedule-specific baseline, explicit missingness accounting and outcome-blind quality-control procedures.

This dissertation thus examines the timing of heart-rate sensing with an explicit consideration of the energy constraint. The central research question is: at a given modelled acquisition schedule-appropriate device-energy budget and with the same modelled device-energy constraints attributed to the sampling schedule and fixed number of alerts, are there benefits with the nocturnally concentrated heart-rate acquisition schedule over the uniform, random-block, and rest-triggered schedules and do these benefits persist over an equal-sample-count daytime-rest control? The term time-of-day effect is used instead of stating a causal circadian mechanism as the public datasets do not allow for sufficient experimental control to separate the endogenous circadian physiology from sleep, posture, behaviour and environment.

The study has four contributions. First, it redefines wearable infection detection as an energy/battery constrained sensing placement problem, not just an algorithm selection problem. Second, it employs a personalised resting-heart-rate detector that does not have a schedule-dependent bias, and has strictly causal baselines and an explicit alert budget. Third, it keeps a deployable nocturnal schedule in mind but distinguishes that from a rest-matched mechanism control, and refrains from overinterpreting a simple night versus day comparison. Fourth, it adds an outcome-blind data-semantics correction that was discovered during quality control, and assaves the failed run as part of the reproducibility record. The core empirical finding is negative but informative: the nocturnal schedule was not superior to the uniform schedule, and this result was robust to the prespecified sensitivity analyses displayed in the core runner.

2. Related Work

2.1. Wearable signals for respiratory-infection detection

Wearable infection detection is based on the concept of the individual as the best control. Before the conventional symptom threshold is reached, acute infection can lead to higher resting heart rates, sleep disturbances, decreased activity or change in temperature. Radin et al. (2020) illustrated the usefulness of this concept at the population level by integrating Fitbit data on resting heart rate and sleep with influenza-like-illness surveillance. Quer et al., (2021) then demonstrated data from wearable sensors could provide discrimination in addition to self-reported symptoms within a population of symptomatic individuals who were being tested for

COVID-19. The studies vary on endpoint and scale but collectively agree with the general principle that longitudinal personal baselines may detect more subtle deviations that is not detected by just one measurement.

A particular study is Mishra et al., 2020 which is relevant because its Stanford Phase 1 data are the basis for the present secondary analysis. The study analyzed 32 COVID-19 positive participants with appropriate wearable data and methods were created such as RHR-Diff and HROS-AD. It noted physiological changes may frequently be found before symptoms occur, and a two-stage, retrospective resting-heart-rate warning system could catch 63% of cases pre-symptomatically. Importantly, its resting-heart-rate construction was based on participant-specific baselines, not a population threshold, which is also done here but not by simple reuse of full-resolution features following schedule masking.

Alavi et al. (2022) was an extension of the Stanford work towards a prospective real-time alerting system. A total of 84 SARS-CoV-2-positive cases were confirmed in 3,318 enrolled participants and provided enough wearable data to signal. Of these 84 cases, 67 (79.8%) resulted in alerts at or before the onset of symptoms and/or diagnosis, with a median pre-symptomatic signal of three days. The work additionally reveals an important practical problem: alerts aren't illness-particular. Physiological anomalies may occur due to stress, travel, alcohol and other factors. In this regard, the current dissertation considers the permitted number of alerts as a part of the operating point and not as a sensitivity measure without a constraint on false alerts.

2.2. Time of day, rest and measurement quality

Physiological and measurement-based motivation to consider time of day. The 24-hour cycle of heart rate has endogenous rhythms as well as rhythms due to changing behaviour. Cardiovascular chronobiology has been reviewed in the context of diurnal fluctuations in heart rate, autonomic regulation, blood pressure and cardiac metabolism (Shea et al., 2026; Zhang et al., 2020). Speed et al. (2023) conducted a large wrist-wearable study and showed that inactive heart rate is lower at night than in the day and that the lowest inactive HR typically happens in the early morning. These observations suggest that nighttime sensing could be interesting since movement artefact and behaviour heterogeneity could be minimized.

However, this should not automatically be considered a circadian intervention if it is conducted at nighttime. Clock time is correlated with sleep, body posture, activity, meals, light and

behavioural routines. A nighttime schedule might thus look better just because of the increased number of resting periods captured. This is done by separating the main comparison of the deployable S3 versus the uniform S2 schedule from M1, which is a comparison of nocturnal S3r versus S6 daytime-rest with matched daily burst opportunity, a mechanism-oriented comparison, referred to as M1. A positive M1 would have strengthened the narrower statement that the time-of-day effect was still present after the matching for gross rest-state; it would not have proven an endogenous circadian effect.

2.3. Observation cost and low-power wearable design

Energy-aware sensing has been a long history in theory. Banerjee and Veeravalli (2012) introduced quickest change detection with on-off observation control, where there was an explicit constraint on the number of observations used before the change and a requirement of low detection delay. The mathematical context is different for consumer wearables, but the design reasoning is applicable: Sampling policies should be evaluated in conjunction with false alarms and detection delay. When observations are context dependent, random fractional sampling is not necessarily the best choice.

2.4. Methodological risks: leakage, missingness and post-event behaviour

If future information is used in the preprocessing, retrospective wearable studies may overestimate performance. There are two risks that are particularly pertinent. First, baseline normalisation should not involve data from the evaluation period or information that is available only in the complete unmasked signal. Secondly, behaviour might alter following diagnosis or testing. Cleary et al (2022) pointed out that behaviour after receipt of a COVID-19 test result could be what is captured in the wearable signal and not necessarily the physiology of infection. This study thus defines a primary window that occurs prior to symptoms (ending at onset minus one day) and excludes any windows that are adjacent to infection from negative calibration, while estimating the baseline within the schedule's remaining samples.

3. Research Gap and Hypotheses

Although there are existing studies of wearables for infection, most of these studies do not control for the effect of sensing placement under a fixed resource envelope while examining longitudinal physiology for early warning. A schedule with more minutes sampled, or with a

less expensive activity-detection mechanism, or with an algorithm that evaluates it based on the period evaluated by the algorithm may look better for non-temporal reasons. The question is thus not whether or not there is a change in HR during infection, but whether the same nominal schedule-attributable energy budget can be smartly spread throughout the day.

This question was operationalized by four preregistered comparisons. The primary confirmatory test was carried out on H1 which compared nocturnally concentrated S3 with uniform-block S2. M1 was a comparison of nocturnal rest-only S3r and daytime-rest S6, and was used to determine if any nocturnal advantage remained following gross rest-state matching. H2 contrasted the S3 with seeded random-block S5 and H3 contrasted S3 with causal rest-triggered S4. The family-wise error rate for the entire procedure was set at 0.05, meaning that H1 was required to reject at two-sided alpha 0.05 with $\theta>0.5$ before M1 could be confirmatorily tested, which in turn had to be tested at two-sided alpha 0.05 with $\theta>0.5$ before H2 and H3 could be tested using Holm-Bonferroni correction.

4. Methodology

4.1. Study design and reproducibility strategy

The design was a staged preregistered, reproducible, retrospective secondary analysis of public wearable data. Wherever possible, methodological choices were made in Git first and scientific results were then checked. The corrected primary experiment was committed to as 83cc8d1, and, much later, post-primary sensitivity operationalization froze at c6c9cb4, with a specific note that it would not re-open confirmatory inference. The last commit to the tables-and-figures code was at 9e9b592. This sequence is crucial in order to differentiate between the decisions that were made in advance and the implementation of the decisions which happen after the results.

4.2. Data sources and cohort construction

The data was obtained from the public archive of the Stanford Phase 1 program (Mishra et al. 2020) and the Stanford Phase 2 program (Alavi et al. 2022). There were 116 positive infection records available for the two releases. Heart-rate and step streams were analyzed with the onset or diagnosis anchors provided by the endpoint. Where the release enabled, device provenance was maintained, for the schedule simulation the primary pipeline of raw data was limited to

compatible Fitbit streams, with Fitbit and Apple Watch cases also included as Phase 2 candidates.

Cohort selection was intentionally broken down into two components: source availability and detector evaluability. 49 participants entered the pre-calibration stage following the onset, device and coverage checks. A total of 38 individuals completed the frozen base; 10 from Phase 1 and 28 from Phase 2. The calibration-availability distribution was not compatible with retaining at least 80% of the participants at the candidate minima above 28 days, with the result that the frozen availability-only rule was used, and C_min was reduced to 28. The number of participants N was permitted to vary pairwise due to a requirement of C_floor=16 z-defined calibration days for each arm. There were N=30 for all four major comparisons in the corrected run.

Table 1: Final cohort and corrected-run quality-control summary

Item	**Value**
Public positive/onset-label records	116
Pre-calibration participants	49
Final frozen base cohort	38 (Phase 1: 10; Phase 2: 28)
Primary pairwise N	30 for H1, M1, H2 and H3
Primary alert-budget violations	0
Participants >40% unavailable M1 event-days	0
Median M1 unavailable fraction	0.0
Corrected Phase 2 sparse-step semantics applied	28/28 Phase 2; 0/10 Phase 1

4.3. Outcome-blind Phase 2 sparse-step correction

In fact, during QC, an unlikely trend was found: rest-dependent schedules provided virtually no data for a large proportion of Phase 2 participants, but clock-driven schedules worked as expected. There was no inspection of scientific win/loss counts, detection rates, warning times, p-values and effect sizes. The answer was revealed by an outcome-blind raw-input diagnostic: Phase 1 step files explicitly included zero-step minutes, but Phase 2 Fitbit step files had very few minutes of steps and lacked the zero-step minutes, even though they had extensive coverage of heart-rates. The same source convention was documented independently by the public Anomaly Detect implementation: that its raw step stream did not include zeroes, but added them based on the heart-rate timestamps if step rows were missing (gireeshkbogu, n.d.).

Amendment A6 consequently specified a minute-level correction just for the Phase 2 Fitbit data. Heart rate was measured in a minute and if there was no explicit step-file record for that minute, then steps in that minute were assumed to be 0. Explicit step records always prevailed,

including the explicit missing value, and the minutes for which there were no steps or heart rate were unknown. Phase 1 was left unchanged. The correction was not done after any scientific results were examined, and the superseded run was not overwritten. The rule was applied to all 28 Phase 2 participants and no one in Phase 1 (10 participants). The median for the number of minutes that had to be inferred to zero in Phase 2 was 214,578.5, verifying that the correction covered a very widespread representation difference and not just a few records.

4.4. Schedule definitions

Seven minute level schedule masks were implemented. The continuous reference stream was called S1. The uniform-block schedule, S2, distributed N periods of 10 minutes each evenly throughout the 24 hours. S3 focused N 10-minute episodes in the designated night time period (00:00-06:59). S5 implemented seeded non-overlapping random blocks, which is a timing-agnostic sparse comparator. S4 was a causal rest-triggered schedule, which means that it only required information on what happened in the preceding minute before acquisition, a short time "rest" hold period, and future activity could not be used to decide earlier sampling.

For M1, two counterfactual analysis constructs were used instead—constructs that were not deployable online schedules: S3r and S6. Both were rest-conditional, and where possible had the same number of matched burst counts during the day, N'. Bursts were permitted only during nocturnal window and at rest for S3r and only within the daytime rest for S6. There were no participant-days available for M1 with too little rest opportunity in either window. The construction thresholds preregistered were N_min=2. Since the daily rest-opportunity pattern was used to create like-for-like daily counts, M1 can be thought of as a mechanism-oriented rest-matched analysis, and not a causal sensing policy. At the corrected run, no participant had more than 40% unavailable event days; the highest percentage of unavailable event days at the participant level was 0.238; M1 was not inconclusive due to construction.

Table 2: Sensing schedules at the primary E3 budget

Arm	Purpose	Placement	Primary E3 allocation
S1	Reference	Continuous observed stream	Continuous
S2	H1 comparator	Equal-spaced 10-min blocks across 24 h	7 x 10 min/day

Arm	Purpose	Placement	Primary E3 allocation
S3	Primary nocturnal	10-min blocks within 00:00-06:59	7 x 10 min/day
S5	H2 comparator	Seeded non-overlapping random blocks	7 x 10 min/day
S4	H3 comparator	Causal rest-triggered blocks	Budget/cap equivalent; may underspend if no trigger
S3r	M1 mechanism arm	Nocturnal rest-only	Daily matched N' with S6
S6	M1 control	Daytime rest-only	Daily matched N' with S3r

4.5. Modelled energy constraint

Physical prototype was not bench-validated in the final design. Thus, all the energy claims are modelled explicitly and not measured. Energy model included sensor wake-up, sensor settling, active acquisition, idle overhead, low-power activity sensing, active accelerometer cost, local evaluation and Bluetooth communication. These components were represented by uncertainty in the form of three parameter scenarios: LOW, CENTRAL and HIGH. The CENTRAL/E3 operating point was the main operating point. The choice of burst count was made a priori to be the burst integer N that minimized the absolute relative error between the modeled energy of N bursts count and the nominal energy target defined by the S1, and in case of a tie, the lower-energy option was chosen. Only schedules within the modelled errors ± 5% were considered to be energy-matched. The relevance of such component-level accounting is demonstrated by modern low-power wearable systems (Dheman et al., 2024).

The ten-minute bursts that were frozen at E3 were the same for all LOW, CENTRAL and HIGH energy parameterizations (N=7). The nominal target under CENTRAL/E3 was 376.889 J/day while clock-driven sparse arms S2, S3 and S5 were modelled at 366.834 J/day (-2.668%) and rest-conditional S4, S3r and S6 were modelled at 372.502 J/day (-1.164%), within the frozen ±5% matching rule. This is helpful because the conclusion at E3 shouldn't be the result of a specific LOW/CENTRAL/HIGH setting on a component. But equal budget does not mean equal realised energy for a reactive schedule: S4 can underspend by triggering less than seven bursts if there are no rest opportunities. The research claim is thus presented as a comparison under the same schedule budget or cap (modelling), not as identical observed battery consumption (measurement). The frozen parameter scenarios and E3 matching audit are included in Appendix C.

4.6. Causal TOD-z detector

The key detector used was a resting heart-rate anomaly score that had been customised to include time-of-day normalisation. After applying each schedule mask, a minute was considered resting when the step count was zero. At least 5 retained rest samples were needed to be baseline-eligible for a day. The baseline for each prediction day d used the latest 28 days prior to d that qualified as baseline days (days with at least five retained rest samples), within a 90 calendar day lookback. This one-pass construction eliminated the use of nested baselines and guaranteed that the baseline estimate for the sparse schedule was obtained from samples that would have been selected by the schedule.

The algorithm used in the selected historical days to estimate a participant-specific mean resting heart rate by hour of day, $\mu(p,h)$ that was used to calculate the daily deviation D. The median and median absolute deviation of the historical D values from this same causal baseline form the robust center and scale for the current day's deviation D. To avoid being divided by very small variability, the scale floor was set at one beat per minute. The z value was thus only achieved when the necessary historical baseline was found. If there was an inadequate history or the data for retained rest was not available, it was specifically noted as being unevaluable rather than as a normal score.

The time-of-day normalisation was selected to avoid a schedule being favoured simply because it had sampled a period of lower or higher heart-rate of the usual day. This normalisation was removed by prespecifying a sensitivity without changing the causal 28-day/90-day structure. Especially this was the case for the research question, that if it was found that S3 was superior only if normal variation in the time of day was disregarded, this would make the result hard to interpret as a proof of usefulness of a sensing schedule.

4.7. Calibration and alert burden

A new threshold τ was freely determined from the history of the negative z-defined days only for each participant-arm. Calibration days were outside all infection exclusion periods and lasted until the onset minus 28 days. The threshold grid was set to every 0.1 from 1.0 to 6.0 and then +infinity for the rest. The smallest finite threshold for which the alert burden was met was chosen; if no finite threshold met the burden, +infinity yielded no alerts. The main alert load was at most 2 alert-days per person-month. Sensitivities were repeated at one and four alert-days per person-month.

The persistence rule used two observations at a high z before entering into an alert state with a maximum gap of 3 days to control continuity. Thresholds were fixed for the main evaluation period. A separate sensitivity recalibrated τ causally on expanding negative history, Importantly, neither the infection outcome nor warning time was used to select thresholds. The primary run was corrected and no alarm-budget violations occurred and no arms were saturated, thus demonstrating that the planned operating point was possible for all pair-eligible participants.

4.8. Endpoint and hierarchical paired comparison

The first outcome was warning before symptom onset, defined as between onset-21 days and onset-1 days. Descriptive at-or-before-onset performance was recorded separately on day zero. If an arm did not have a z-defined day at any time in the pre-symptomatic window, that arm was not available for that participant comparison and not counted as failed detection. When it's assessable, an alert before the symptoms was counted as detection and warning time was the number of days between the earliest qualifying alert and the onset.

The paired endpoint was the combination of detection and timing, but did not provide an arbitrary warning time for non-detections. In a comparison A versus B, A was deemed to be the winner if A detected pre symptomatically and B did not, or if both detected and A warned earlier. B was the winner for the symmetric conditions. There was a "Tie" if neither detected or both detected with the same warning time. The effect estimate was $\theta = (W+0.5T)/N$, where ties were split evenly, and interpreted as the probability that a randomly selected paired participant favors A over B. Therefore, a value of $\theta>0.5$ is in favor of A and a value of $\theta<0.5$ is in favor of B, and $\theta =0.5$ represents no paired advantage.

4.9. Statistical analysis and sensitivity plan

The two sided binomial sign-test p-value, based on discordant W and L pairs conditional on W+L, was therefore not affected by ties although θ was. 10,000-resample participant-level BCa bootstrap 95 per cent confidence interval for θ. The serial gatekeeping procedure was used with H1 being tested first at $\alpha=0.05$, and then the M1 gate being opened when H1 was rejected and when $\theta>0.5$. If M1 also rejected in positive direction then H2 and H3 would be tested with family alpha = 0.05 based on Holm-Bonferroni. No formal equivalence margin was prestated, so that a failure to reject H1 does not imply that schedules are equivalent.

All sensitivity analyses were in nature estimation only and did not reopen the confirmatory gate. The following core analyses were performed: Phase 2-only cohort, a common intersection cohort meeting C_floor in all primary arms, finer calibration resolution, no multiple episode rows, alert burdens B=1 and B=4, no time-of-day normalisation, rolling threshold recalibration and realized-minute matching for M1. The M1 count-matching algorithm was only used after the primary result was known, and was therefore openly reported as post-primary: Each participant day was thinned down symmetrically (not using heart-rate values or outcomes) to the lower of two arms delivered HR-observed minute counts.

5. Results

5.1. Corrected-run quality control

The corrected run comprised the anticipated 266 arm-level records (38 x 7 arms). The 28 of 28 Phase 2 participants received Phase 2 sparse-step semantics, while the 10 Phase 1 participants did not receive sparse-step semantics. Calibration-floor attainment was 32/38 for S1, S2, S5 and S4, and 30/38 for S3, S3r and S6. 6 participants had no evaluable pre-symptomatic days in S1, S2, S5 and S4, and 8 participants in S3, S3r and S6. Excluding these unavailable cases pairwise, not as negative detections, resulted in N=30 in each confirmatory comparison.

The delivery was as planned since the A6 correction. The median number of scheduled minutes per sparse arm was 1,470 (7 10-minute bursts a day) throughout the 21-day event window. Median delivered minutes were 1,386 for S2, 1,437.5 for S3, 1,399.5 for S5, 1,458.5 for S4, 1,403 for S3r and 1,396.5 for S6. For all arms, the median number of event days with some input data were 21 event days. No row was scheduled more than their requested count, no delivered-minute count surpassed their scheduled minutes and no alert-budget violations occurred.

5.2. Primary confirmatory result: H1

H1: Nocturnally concentrated S3 vs uniform-block S2. In 30 paired participants S3 won 8, S2 won 12 and there were 10 ties. The paired effect was $\theta=0.433$ (95% BCa CI 0.300-0.583), with an exact two-sided sign-test p-value of 0.5034. Thus, the directional gate condition ($\theta>0.5$) was not fulfilled and H1 was not rejected. The fixed-sequence confirmatory gate was closed at stage 1.

Descriptive results of the arm-level were consistent with the paired endpoint. Pre-symptomatic detection rate was 63.3% (19/30 evaluable participants), and the median warning time for the 19 pre-symptomatic cases detected by S3 was 13.0 days (IQR 9.0-19.5). S2 detected 20 of 30 (66.7%), with median warning of 14.0 days (IQR 8.2-19.0). Although these are not independent arm tests as the main analysis was paired, these summaries indicate that it was not a single part of the hierarchy that led to no S3 advantage.

Table 3: Primary paired results at CENTRAL/E3

Test	Comparison	N	W/L/T	θ (95% BCa CI)	Exact p	Detection A vs B	Status
H1	S3 vs S2	30	8/12/10	0.433 (0.300-0.583)	0.5034	63.3% vs 66.7%	Not rejected
M1	S3r vs S6	30	13/11/6	0.533 (0.367-0.683)	Not tested*	76.7% vs 63.3%	Gate closed; estimation only
H2	S3 vs S5	30	10/10/10	0.500 (0.350-0.650)	Not tested*	63.3% vs 63.3%	Gate closed; estimation only
H3	S3 vs S4	30	2/9/19	0.383 (0.283-0.483)	Not tested*	63.3% vs 73.3%	Gate closed; estimation only

* M1, H2 and H3 were not confirmatorily tested because H1 did not open the fixed-sequence gate. Their raw descriptive sign-test values remain in the audit outputs, but they are not reported here as confirmatory p-values.

5.3. Estimation-only later comparisons

M1 could not be tested confirmatorily but its estimate was more or less neutral. S3r versus S6 produced W/L/T=13/11/6 and θ=0.533 (95% BCa CI 0.367-0.683). 23/30 (76.7%) participants were detected by S3r, and 19/30 (63.3%) by S6; the median warning among detections was 13 days for S3r and 9 days for S6. A time-of-day mechanism cannot be supported by this pattern due to the gap in the interval, the gate closure and instability of the fine-calibration subset.

H2, S3 versus random-block S5, was exactly balanced on the paired effect: W/L/T=10/10/10 and θ=0.500 (95% BCa CI 0.350-0.650). The sensitivity of both arms was 63.33% (19/30 participants). There was a trend towards a more marked descriptive pattern for the rest-triggered comparator with H3: S3 versus S4 yielded W/L/T=2/9/19 and θ=0.383 (95% BCa CI 0.283-0.483). S4 detected 22/30 participants (73.3%) and had a median warning of 15.5 days, compared with 19/30 and 13.0 days for S3. This is an exploratory observation and not evidence of superiority as the confirmatory gate had closed.

The H3 result is also a good example of that fact that the preregistered test and effect interval should never be confused. The BCa interval for θ is below 0.5, but the actual sign-test p value on discordant pairs equals 0.0654, based on the eleven non-tied pairs (two wins and nine losses for S3). There are 19 ties for which the weight is ½, both within the computation of θ, and within the computation of the bootstrap distribution of θ. The sign test was designated as the confirmatory decision rule in the preregistration and, in any event, was never reached by the gate.

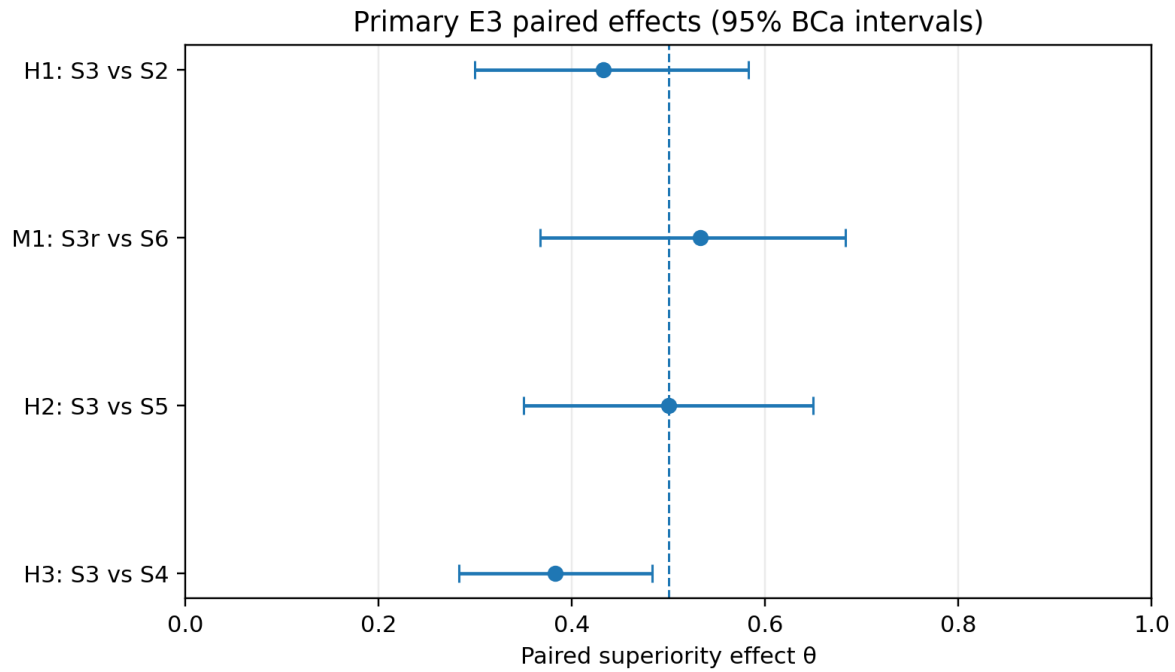


Figure 1: Primary paired superiority effects. The dashed line at θ=0.5 denotes no paired advantage. M1, H2 and H3 are estimation-only because H1 did not open the fixed-sequence gate

5.4. Sensitivity analyses

The main H1 conclusion remained unchanged in all main core sensitivity analyses. In the Phase 2-only cohort, θ was 0.420 (95% BCa CI 0.260-0.580). The primary result was exactly reproduced at θ=0.433 using the common-cohort restriction. Finer calibration resolution was used to restrict N to 18, and θ to 0.444 (0.250-0.639). This was rerun without multiple-episode rows and the same primary result was obtained. The value of θ was 0.483 at an alert budget of one alert-day per person-month and 0.450 at an alert budget of four alert-days per person-month. Without time-of-day normalisation, it was θ=0.450, and with rolling threshold recalibration, it was θ=0.383. Hence, none of the sensitivity estimates exceeded 0.5 for H1.

The sensitivity for the M1 realized-minute matched was particularly valuable for the information on source coverage. The original M1 resulted in a W/L/T of 13/11/6 and a θ of 0.533, while M1 after deterministic per-day thinning to equal delivered HR-observed minute counts produced the same W/L/T and θ of 0.533; a 95% BCa interval of 0.367-0.700. This indicates that the point estimate for M1, described above, may not have been due to the small pre-masking advantage seen at night. However, the fine calibration subset pushed M1 to θ=0.389, thus helping to maintain the non-aggression stance towards the mechanism estimate.

The stability of H3 was lower in the operating points. The majority of sensitivities remained descriptively in favour of S4 with θ around 0.38-0.42 for S3 and S4 and the stricter B=1 alert budget gave a value of θ=0.517. The sign change is indicative of the fact that any apparent benefit of reactive rest-triggered sampling is dependent on the operating point of the alert and is not a stable secondary finding. H2 did not vary much from neutral in all the sensitivity analyses reported.

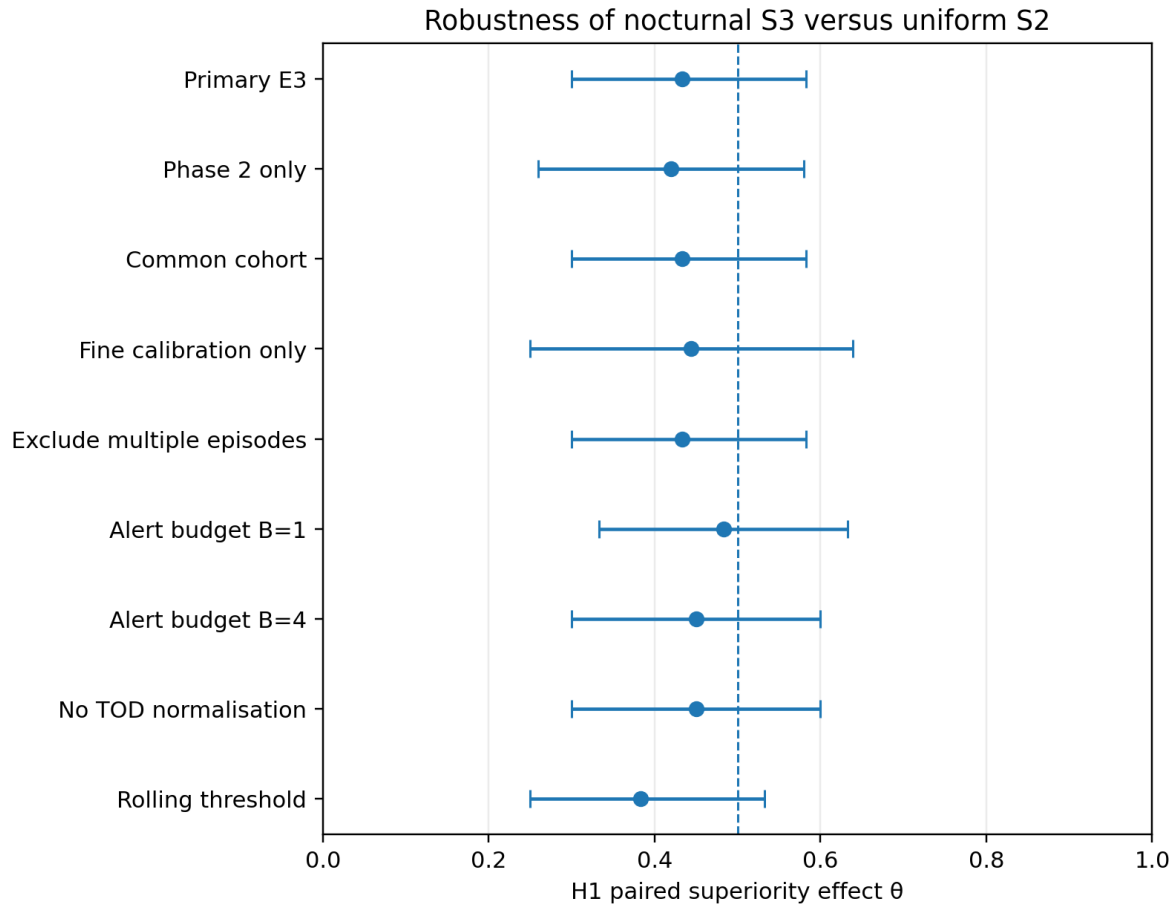


Figure 2: H1 robustness across core sensitivity analyses. All reported point estimates remain at or below

ϑ=0.5.

6. Discussion

6.1 Principal finding

Their main conclusion is simple: When the same number (7) of 10 minute heart-rate acquisition bursts were concentrated in the nocturnal window (versus uniformly distributed throughout the day) at the energy budgets tested (E3), there was no improvement in pre-symptomatic warning. In fact the paired estimate numerically favoured S2 although the confidence interval was wide and the exact Sign test was far from being significant. This conclusion is reinforced by the sensitivity analyses which did not yield an H1 point estimate greater than 0.5. It is thus a 'no evidence of nocturnal superiority' rather than 'proof of equivalence'. The 95% interval still contains numbers greater than 0.5 and the sample of 30 paired participants is not large enough to exclude modest advantages or disadvantages with large degrees of confidence.

This result is important as the biological intuition for night-time sensing is strong. When asleep, resting heart rate is less clogged and generally lower, and can be measured during early morning to obtain a baseline reading. So, if detection were just based on that intuition then one could have thought that S3 would have dominated S2. The uniform schedule, however, had a slightly higher rate of detection and warned one day earlier than the median duration of warning for cases that were detected. One possible explanation is that the deviations to normal physiology due to infections are not limited to night time physiology; spreading the sampling over the whole day might allow a more complete set of abnormal periods to be identified and the time-of-day normalisation of the detector accounts for much of the normal diurnal/ circadian structure anyway.

6.2. Why a nocturnal advantage may not have emerged

There are several ways to obtain the observed null. First, the main detector specifically standardised resting heart rate with regard to the hour of the day. This was a conscious fairness functionality, however it additionally signifies that S3 could not gain benefit by sampling a naturally low heartbeat window due to time of day normalisation. Without time of day normalisation, the sensitivity was still not in favour of S3 and this design decision was not able to remove a significant hidden advantage. Second, the uniform schedule they employed was blocks and not isolated 1 minute samples. Seven 10-minute bursts over 24 hours still allow opportunities to meet rest – S2 isn't a poor base line in this regard. Third, if the infection signal

may be persistent over the course of the day, the placement is not as critical as getting adequate clean rest observations to keep a causal baseline.

Fourth, a higher concentration during the night decreases temporal diversity. Changes in behavioural routine, recovery from activity and daytime inactivity could be under-sampled by S3, as it is during these periods that a physiological deviation may be most strongly present. The study does not answer this question as to which of these explanations is most prevalent, but it does call into question an assumption that ‘cleaner' night-time measurements necessarily maximise earliest warning. This implies that for device designers, to reduce the sensing to a desired physiological window is not necessarily optimal and must be verified as such with a simple uniform duty cycle.

6.3. Interpretation of the rest-matched mechanism analysis

M1 was developed to differentiate time of day from gross rest state matching, and due to H1's failure it was not possible to confirm via testing. The descriptive point estimate of $\theta=0.533$ for S3r versus S6 is close to neutral and its interval is wide. The realised-minute-matched sensitivity achieved the same wins, losses and ties, which is reassuring with regard to the small difference in source-coverage between night and day. But the fine-calibration subset moved the point estimate below 0.5. These are counterfactual rest-matched arms and are not causal schedules that can be deployed. Overall, the data do not suggest that there is special predictive information in nocturnal rest, over and above daytime rest.

It's a distinction that is important for terminology. Although, it would be tempting to say that the study describes a 'circadian schedule' although the internal circadian phase was not manipulated or estimated. The intertwining of clock time and sleep, rest and behaviour continues. Even if the M1 had been positive, the time of day association would have been possible only using the gross-rest matching. The current findings are even weaker: The mechanism gate never opened. In the Results, causal circadian terms are thus avoided and the circadian physiology is only introduced as background and as a potential mechanism.

6.4. Rest-triggered sensing as an exploratory direction

S4 is the most interesting exploratory comparator. Based on the primary E3 estimate, rest triggered S4 exhibited a higher detection and longer median warning than S3, and $\theta=0.383$ for the comparison of S3 versus S4. This descriptive pattern does not imply that context-aware

sampling is necessarily better than fixed clock-time concentration; it simply suggests that it is a motivating pattern for future research into context-aware sampling. A hypothesis like this is consistent with observation-control theory: When the samples are more informative in a known state (e.g., rest), a causal policy that waits for this state can be efficient in using a limited acquisition budget.

But, don't over sell the H3 pattern. The fixed-sequence gate was used to prevent confirmatory testing, the number of discordant pairs was small, and the B=1 sensitivity caused the point estimate to be slightly towards s3. S4 also incurs a modelled activity-sensing energy cost that is based on the activity being used, as well as a chance to underspend the configured budget, in the event that rest triggers are unavailable. It should be noted that this simulation should not be used to extrapolate between the context-triggered and clock-driven policies and that a future prospective device study entails measuring the current consumption and using a predeclared battery-energy budget.

6.5. Implications for low-cost wearable design

The bottom line: simplicity of scheduling might be worthwhile. The use of uniform duty cycling is simple to implement and is predictable for battery budgeting and in this cohort was competitive with a slightly better descriptive detection and median warning compared to the use of nocturnal concentration. A designer, therefore, who wants to minimize the photoplethysmography duty cycle may wish to design from a uniform-block baseline and require evidence before adding the time-window specific logic. The highlight of this E3 result is also a clue to the energy modelling and physiological performance separation. The study cannot claim how many days of battery life a real device would gain because hardware energy was not measured. All three parameter scenarios (LOW, CENTRAL and HIGH) mapped E3 to the same seven 10-minute bursts; therefore, the physiological conclusions remain stable to these uncertainties in the parameters.

6.6. Methodological lessons from the A6 correction

An important reproducibility lesson, the Phase 2 sparse-step correction. The first run went through a lot of structural tests and yielded implausible behaviour in rest-dependent arms. If the scientific outcomes had been examined prior to the rationalisation then there could be the danger of rationalising it afterwards. Instead, a diagnostic was outcome-blind, the source semantics were compared to the public Stanford code, the correction was noted in a new

amendment and the failed run was kept. This isn't the end of researcher freedom, but it does provide a definite audit trail that would differentiate an error in a representation of the raw data from model tuning to get the result they wanted.

7. Limitations

One of these restrictions is sample size. The confirmatory paired analysis evaluation included 30 participants of the 116 identified in the public labels, after source, baseline and arm-specific evaluability criteria. This is the proper denominator to use for inference, but this results in wide intervals. A moderate disadvantage can't be excluded, nor can a modest advantage, by the H1 interval of 0.300-0.583. This study thus had a greater likelihood of identifying a large rather than a subtle scheduling effect.

Second, this is not a prospective wearable schedule simulation but is a retrospective one. Masking was done for the data already recorded and scheduled. It is possible to estimate what measurements would have been kept under each schedule, but the modelling of hardware interactions like sensor thermal behaviour, acquisition failure due to a moving sensor, firmware latency, packet loss, as well as user charging behaviour and battery-voltage effects are not considered. Energy was modelled by means of component terms and uncertainty scenarios were modelled; no current traces or battery-discharge experiments were collected. The claims shall thus be restricted to energy modelled with a schedule.

Thirdly, the public releases were diverse. There were differences in the number of raw steps represented in Phase 1 and in Phase 2, and it was not possible to confidently assess release overlap due to the lack of consistency in de-identified archive dates and with published study windows. The main conclusion was not reliant on the cross-release heterogeneity, which led to reduced concern for the Phase 2-only sensitivity, but it cannot prove independence of participants. There is also imperfect control over the effect on the devices, since the releases aren't creating a uniform per-participant device map for all the users of the historical archives.

Fourth, a fixed window of the night (00:00-06:59), instead of personalized sleep or biological night, is used for the analysis. There is a variation in chronotype, working schedule, sleep time etc. A participant that sleeps between 22:00 and 05:00 is considered the same as a participant that sleeps between 03:00 and 10:00. Sleep-record-defined Phase 1 analysis was preregistered as exploratory but the primary question across cohorts could not rely on the sleep records which were not available in Phase 2.

Fifth: The alert detector is by design simple and targeted. TOD-z was not selected because it is not necessarily the most accurate infection detector, but because it has a causal transparency and schedule neutral. Potentially, they could be multivariate, using other features like temperature, sleep architecture, heart-rate variability or respiratory rate. But in this small cohort there would be increased training, leakage and generalisation threat with increased model complexity. Thus, the current results are about the interplay between the sensing schedule and this type of resting-heart-rate anomaly detection and not all machine-learning models.

Sixthly, symptom onset and diagnosis anchors are imperfect. There may be some variability in self-reported onset dates and asymptomatic cases need a diagnosis anchor not a symptom anchor. Also, the primary window uses the assumption that a warning in the 21 days before onset may be of relevance. Very early alerts could be due to other stresses and not infection. This ambiguity is lessened but not eliminated by the fixed burden of alert, as anomalies on wearables do not necessarily translate into pathogens.

Lastly, the negative outcome is not supposed to be changed into an equivalence statement. There was no pre-registration of a formal equivalence margin and the confidence intervals are still large. The study demonstrates that the tested design does not support nocturnal superiority and that all schedules aren't equivalent and that seven 10-minute bursts cannot be said that it is globally optimal. The values and rankings may vary with different duty cycles, richer sensors, or with different alert burdens or populations.

8. Reproducibility, Data Governance and Ethics

The analysis was based on de-identified public data sets provided by the original Stanford studies where participant recruitment and data collection were done under the respective institutional review processes. No new participants were recruited, no attempts were made to contact anyone, or to re-identify participants in this dissertation. Only data fields were processed that are necessary for simulating the schedule, onset anchoring and reproducibility. Sensitive raw archives should be kept access controlled even if the public version of the archive is de-identified in any shared project space.

Version controlled code, frozen amendments, automated tests and immutable result directories were used to ensure reproducibility. The project test suite started with 154 passing tests and 5 more tests were added in the post-primary sensitivity operationalisation, resulting in 159

passing tests before the sensitivity run. The main corrected run was related to Git commit 83cc8d1, the sensitivity plan (table) to c6c9cb4 and the final table figure generator to 9e9b592. Four upstream repositories were pinned for the reproduction audit (mwgrassgreen, n.d.-a, n.d.-b; gireeshkbogu, n.d; StanfordBioinformatics, n.d). Locally, the superseded pre-correction results directory was not deleted. A standalone confirmatory-summary script (reveal_primary_e3_confirmatory_results_fixed.py) was also kept with the audit materials and should be included with any release in a public repository to allow for the final summary CSV and gatekeeping JSON to be re-generated directly.

9. Conclusion

This dissertation compared the effectiveness of sparse heart-rate sensing at different times in the same modelled energy budget/alert burden for the purpose of pre-symptomatic warning of infection. In the primary comparison, seven 10-minute bursts per day (concentrated during the night) were not more effective than seven uniformly distributed bursts. The paired superiority estimate was θ=0.433(95% BCa CI 0.300-0.583), p=0.5034 and all core estimates of H1 sensitivity were ≤ 0.5. Thus, the pre registered gate shut earlier than the confirmatory test of the mechanism and schedule comparisons.

It's not like timing is not important. Instead, there was no intrinsic support for the simple assumption that night time concentration is necessarily improved in this cohort. The descriptive rest-triggered pattern is hypothesis-generating and uniform duty cycling was a competitive baseline . Future studies should confirm this comparison on a larger, independent sample of participants, incorporate the use of participant-specific sleep or circadian phase (when available), incorporate actual energy measurements from the hardware, and test context-aware policies in the context of real battery constraints. More generally, wearable early-warning studies should consider sensing policy, missingness, calibration and alert burden as an integral component of the predictive system as opposed to a pre-processing functionality.

References

[1] Alavi, A., Bogu, G. K., Wang, M., Rangan, E. S., Brooks, A. W., Wang, Q., Higgs, E., Celli, A., Mishra, T., et al. (2022). Real-time alerting system for COVID-19 and other stress events using wearable data. Nature Medicine, 28, 175–184. https://doi.org/10.1038/s41591-021-01593-2

[2] Banerjee, T., & Veeravalli, V. V. (2012). Data-efficient quickest change detection with on-off observation control. Sequential Analysis, 31(1), 40–77. https://doi.org/10.1080/07474946.2012.651981

[3] Cleary, J. L., Fang, Y., Sen, S., & Wu, Z. (2022). A caveat to using wearable sensor data for COVID-19 detection: The role of behavioural change after receipt of test results. PLOS ONE, 17(12), Article e0277350. https://doi.org/10.1371/journal.pone.0277350

[4] Dheman, K., Giordano, M., Thomas, C., Schilk, P., & Magno, M. (2024). i-CardiAx: Wearable IoT-driven system for early sepsis detection through long-term vital sign monitoring. In Proceedings of the 2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI) (pp. 97–109). https://doi.org/10.1109/IoTDI61053.2024.00013

[5] gireeshkbogu. (n.d.). AnomalyDetect [GitHub repository]. https://github.com/gireeshkbogu/AnomalyDetect

[6] Mishra, T., Wang, M., Metwally, A. A., Bogu, G. K., Brooks, A. W., Bahmani, A., Alavi, A., et al. (2020). Pre-symptomatic detection of COVID-19 from smartwatch data. Nature Biomedical Engineering, 4, 1208–1220. https://doi.org/10.1038/s41551-020-00640-6

[7] mwgrassgreen. (n.d.-a). Alarm [GitHub repository]. https://github.com/mwgrassgreen/Alarm

[8] mwgrassgreen. (n.d.-b). WearableDetection [GitHub repository]. https://github.com/mwgrassgreen/WearableDetection

[9] Quer, G., Radin, J. M., Gadaleta, M., Baca-Motes, K., Ariniello, L., Ramos, E., Kheterpal, V., Topol, E. J., & Steinhubl, S. R. (2021). Wearable sensor data and self-reported symptoms for COVID-19 detection. Nature Medicine, 27, 73–77. https://doi.org/10.1038/s41591-020-1123-x

[10] Radin, J. M., Wineinger, N. E., Topol, E. J., & Steinhubl, S. R. (2020). Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: A population-based study. The Lancet Digital Health, 2(2), e85–e93. https://doi.org/10.1016/S2589-7500(19)30222-5

[11] Shea, S. A., Scheer, F. A. J. L., Gumz, M. L., Eikenberry, S. A., Qian, J., Thosar, S. S., Sole, M. J., & Martino, T. A. (2026). Unlocking the potential of circadian biology for cardiovascular health. Physiological Reviews, 106(3), 1195–1262. https://doi.org/10.1152/physrev.00015.2025

[12] Speed, C., Arneil, T., Harle, R., Wilson, A., Karthikesalingam, A., McConnell, M., & Phillips, J. (2023). Measure by measure: Resting heart rate across the 24-hour cycle. PLOS Digital Health, 2(4), Article e0000236. https://doi.org/10.1371/journal.pdig.0000236

[13] StanfordBioinformatics. (n.d.). Wearable-infection [GitHub repository]. https://github.com/StanfordBioinformatics/wearable-infection

[14] Tognotti, A., Otesteanu, C. F., Anceschi, L., & Menon, C. (2026). Baseline normalisation choices inflate classification performance in wearable health monitoring: Quantification and mitigation strategies. Frontiers in Digital Health, 8, Article 1827279. https://doi.org/10.3389/fdgth.2026.1827279

[15] Wallner, J., Berbuir, S., Birner, L., Dendorfer, A., Panthi, B., Rahnsch, B., Muma, J., Munga, S., Obor, D., Bärnighausen, T., et al. (2026). Overcoming data loss in wearable disease detection with GAN-based imputation. npj Digital Medicine, 9, Article 275. https://doi.org/10.1038/s41746-026-02518-4

[16] Zhang, J., Chatham, J. C., & Young, M. E. (2020). Circadian regulation of cardiac physiology: Rhythms that keep the heart beating. Annual Review of Physiology, 82, 79–101. https://doi.org/10.1146/annurev-physiol-020518-114349

Appendix A. Full Core Sensitivity Summary

All sensitivity estimates below are estimation-only. No sensitivity p-values were used, and none reopened the confirmatory sequence after H1 failed.

Sensitivity	Comparison	N	W/L/T	θ (95% BCa CI)
Phase 2 only	H1	25	7/11/7	.420 (.260-.580)
Phase 2 only	M1	25	11/9/5	.540 (.360-.720)
Phase 2 only	H2	25	8/10/7	.460 (.300-.640)
Phase 2 only	H3	25	2/8/15	.380 (.260-.500)
Common cohort	H1	30	8/12/10	.433 (.300-.583)
Common cohort	M1	30	13/11/6	.533 (.367-.683)
Common cohort	H2	30	10/10/10	.500 (.350-.650)
Common cohort	H3	30	2/9/19	.383 (.283-.483)
Fine calibration only	H1	18	5/7/6	.444 (.250-.639)
Fine calibration only	M1	18	5/9/4	.389 (.194-.611)
Fine calibration only	H2	18	6/6/6	.500 (.306-.694)
Fine calibration only	H3	18	2/5/11	.417 (.278-.556)
Exclude multiple episodes	H1	30	8/12/10	.433 (.300-.583)
Exclude multiple episodes	M1	30	13/11/6	.533 (.367-.683)
Exclude multiple episodes	H2	30	10/10/10	.500 (.350-.650)
Exclude multiple episodes	H3	30	2/9/19	.383 (.283-.483)
Alert budget B=1	H1	30	9/10/11	.483 (.333-.633)
Alert budget B=1	M1	30	12/9/9	.550 (.400-.700)
Alert budget B=1	H2	30	9/10/11	.483 (.350-.633)
Alert budget B=1	H3	30	6/5/19	.517 (.417-.617)
Alert budget B=4	H1	30	9/12/9	.450 (.300-.600)
Alert budget B=4	M1	30	14/10/6	.567 (.400-.717)
Alert budget B=4	H2	30	11/11/8	.500 (.333-.650)
Alert budget B=4	H3	30	1/8/21	.383 (.283-.467)
No TOD normalisation	H1	30	9/12/9	.450 (.300-.600)
No TOD normalisation	M1	30	13/10/7	.550 (.383-.700)
No TOD normalisation	H2	30	10/13/7	.450 (.300-.617)
No TOD normalisation	H3	30	3/10/17	.383 (.267-.500)
Rolling threshold	H1	30	6/13/11	.383 (.250-.533)
Rolling threshold	M1	30	12/12/6	.500 (.333-.667)
Rolling threshold	H2	30	8/10/12	.467 (.333-.600)
Rolling threshold	H3	30	2/9/19	.383 (.283-.483)
Realised-minute matched	M1	30	13/11/6	.533 (.367-.700)

Appendix B. Audit Trail

Key frozen commits: 990f97e (pre-outcome data/reproduction checkpoint); 4023717 (modelled energy parameters/budgets); d1d5265 (first primary E3 runner, superseded after outcome-blind QC); 83cc8d1 (Phase 2 sparse-step correction and authoritative primary run); c6c9cb4 (post-primary sensitivity operationalisation); 9e9b592 (final results table/figure generator). Any public code release should also include the standalone confirmatory-summary

script retained with the audit materials, together with the pinned upstream commit list and source-archive checksums.

Appendix C. Frozen Modelled-Energy Parameters and E3 Matching Audit

Energy values in this dissertation are engineering model inputs, not bench-measured device consumption. LOW and HIGH are sensitivity scenarios rather than statistical confidence limits. Manufacturer-derived current terms were combined with explicitly labelled engineering assumptions for wake, settling, MCU evaluation and BLE-session duration. No battery-life or measured-current claim is made from these values.

Table 4: Frozen component-level model parameters

Parameter	LOW	CENTRAL	HIGH
E_wake (J)	0.000660	0.008415	0.051150
E_settle (J)	0.001378	0.021810	0.107010
P_active_increment (W)	0.06891	0.08724	0.10701
P_platform_idle (W)	0.00005076	0.00005076	0.00005076
P_accel_lp (W)	0.000033	0.000066	0.000231
P_accel_active (W)	0.00165	0.00165	0.00165
E_mcu_eval (J)	0.000033	0.0001683	0.001023
E_BLE (J)	0.02145	0.2145	2.145

The corresponding engineering assumptions were: wake duration 10/100/500 ms (LOW/CENTRAL/HIGH), PPG settle duration 20/250/1000 ms, MCU evaluation 0.5/2/10 ms, and BLE transmission-session duration 50/500/5000 ms. These durations were prespecified uncertainty settings, not manufacturer specifications.

Table 5: E3 modelled-energy matching audit (seven 10-minute bursts/day)

Scenario	Target (J/day)	Arm family	Achieved (J/day)	Relative error	Status
LOW	297.692	S2/S3/S5	289.458	-2.766%	Within ±5%
LOW	297.692	S4/S3r/S6	292.218	-1.839%	Within ±5%
CENTRAL	376.889	S2/S3/S5	366.834	-2.668%	Within ±5%
CENTRAL	376.889	S4/S3r/S6	372.502	-1.164%	Within ±5%
HIGH	462.398	S2/S3/S5	452.694	-2.099%	Within ±5%
HIGH	462.398	S4/S3r/S6	473.155	+2.326%	Within ±5%

All three E3 scenarios selected N=7 and satisfied the frozen ±5% modelled-energy criterion. Because rest-triggered S4 can fail to trigger all configured bursts, realised energy may be

lower than its configured modelled cap. The audit therefore supports an energy-budget/cap comparison, not a claim of identical measured battery consumption.