Wearable Missingness as a Behavioral Signal: Detecting Wear Pattern Anomalies Around Cardiovascular Visits

Hongzhou Luan [1 2] **Leeor Hershkovich** [3] **Ida Sim** [2] **Jessilyn Dunn** [3]

Abstract

Missing data in wearable device streams is typically treated as noise and is imputed or discarded. We argue that when and how a patient stops wearing their device carries clinically meaningful behavioral information, and propose a personalized anomaly detection framework that models each individual's day-type-specific wear patterns and flags deviations. Applying this framework to Fitbit data from 390 participants (756 cardiovascular emergency visits) in the NIH All of Us Research Program linked with electronic health record (EHR) cardiovascular emergency visits, we find that wear pattern anomalies rise to 1.45x baseline in the week before a visit and to 2.00x baseline in the week after. Critically, aggregate missingness rates show no statistically significant change between the baseline and around visits, indicating that the anomaly detector captures temporal pattern shifts that summary statistics miss.

Introduction

Consumer wearable devices are described as continuously collecting physiological data, yet wearable data inevitably include periods of missing data from charging the device's battery or other non-wear. In research and clinical applications, these gaps are typically handled through imputation or exclusion, treating them as a nuisance to analysis rather than informative observations (Nijman et al., 2022; Lederer et al., 2023). These approaches assume that wearable gaps are missing completely at random (MCAR) or missing at random (MAR), and therefore are not correlated with the user's physiology or behavior. While prior work has noted that wearable adherence varies with health status and demographics (Hershkovich et al., 2025; Dhingra et al., 2023; Hurwitz et al., 2026), no existing framework directly models the temporal structure of non-wear as a behavioral signal for acute clinical event detection.

The MCAR/MAR assumption is naive at best and, at worst, can introduce systematic biases to the data. The decision to remove a wearable device, to alter one's wearing routine, or to abandon the device entirely reflects behavioral patterns that may be influenced by health status, psychological state, or changes in daily functioning (Hurwitz et al., 2026; Attig & Franke, 2020). If these behaviors shift systematically around clinical events, the missingness pattern itself becomes a signal worth modeling. Under this view, where biosignals that might indicate physiological abnormality during acute health events are missed, some wearable gaps are missing not at random (MNAR), where the acute event is tied to the decision to wear or not wear the device. In this case, the deviation from a person's normal routine of wear/non-wear, the missingness mechanism itself, potentially carries information about the very outcomes we seek to predict.

We propose a personalized anomaly detection framework that treats device wear patterns as a structured time series and identifies deviations from each individual's established routine. Rather than analyzing what the device measured, we analyze when during the day the device was worn or not worn, and whether that pattern is unusual for this person. We demonstrate the efficacy of our approach on clinical events extracted from linked electronic health records, enabling a visit-centered evaluation. We apply this framework to Fitbit heart rate data from the NIH All of Us Research Program, linked with cardiovascular emergency department (ED) visit records. Our contributions are:

1. A personalized, day-type-aware anomaly detection framework that treats wearable missingness as a signal.
2. Evidence that wear pattern anomalies cluster around cardiovascular visits in a large, diverse national cohort, with the signal emerging before the clinical encounter.
3. A demonstration that aggregate missingness rates fail to capture this signal, highlighting the importance of modeling temporal patterns of non-wear rather than overall wear quantity.

[1]University of California, Berkeley [2]University of California, San Francisco [3]Duke University. Correspondence to: Jessilyn Dunn <jessilyn.dunn@duke.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

Methods

Data

We investigated Fitbit wearable data in the NIH All of Us dataset (Denny et al., 2019), which is linked to participant EHR and demographic data. The Fitbit data in All of Us is entirely free-living data, meaning device wear was not incentivized and represents the choice of the participant to wear or not wear their device. We include participants who have both Fitbit data and at least one cardiovascular ED visit between January 1st 2017 and December 31st 2024. Each 15-minute slot is summarized into a binary flag of wear (heart rate data present) or non-wear (heart rate data absent), producing a binary matrix of size 96 per day per participant, as in (Hershkovich et al., 2025). Critically, days with no Fitbit records are preserved as all-zero rows rather than discarded, representing complete device non-wear as an informative observation.

We assembled the cohort from the All of Us Controlled Tier (v8) dataset. Eligible participants had (1) Fitbit device data available, (2) completed the Healthcare Access and Utilization survey (concept_id 43528895), and (3) at least one recorded encounter with a visit concept of Emergency Room - Hospital (8870), Emergency Room and Inpatient Visit (262), Emergency Room Visit (9203), or Urgent Care Facility (8782). For these participants, we extracted all condition_occurrence records whose condition_concept_id fell under Cardiovascular finding (SNOMED, concept_id 4023995) in the SNOMED hierarchy, which expands to all specific cardiovascular diagnoses beneath this ancestor concept. We then joined conditions to visits on visit_occurrence_id and retained only visits with at least one cardiovascular condition recorded, yielding a dataset of acute-care encounters linked to cardiovascular diagnoses among Fitbit-wearing participants with healthcare access survey data.

Binary flags of Fitbit wear were analyzed 270 days before and 30 days after each visit date, ensuring sufficient data for establishing baseline device-wear habits and post-visit evaluation. As a lenient inclusion criterion, visits were included if there was any heart rate measurement on at least 30 unique days in the 270 days preceding the visit. Of 2,464 cardiovascular ED visits across 540 participants, 852 visits met this criterion. After further requiring 210 clean baseline days following exclusion of perivisit windows of other visits (see Baseline Construction), the final analytic sample comprised 756 visits across 390 participants.

Baseline Construction

For each participant-visit, we construct a personalized baseline from the 210 calendar days preceding the 60-day previsit test window. To prevent contamination from wear changes potentially associated with other clinical encounters, we exclude any day falling within the perivisit window (-60 to $+30$ days) of any other visit; the 210-day baseline is then drawn from the most recent eligible days. Participant-visits with fewer than 210 eligible days were excluded from analysis. Wearable device wear patterns exhibit strong time-of-day and day-of-week effects (Hershkovich et al., 2025), so we group eligible days into four weekday types $w \in \{\text{Mon}, \text{mid-week}, \text{Fri}, \text{weekend}\}$, where mid-week spans Tuesday–Thursday and weekend spans Saturday–Sunday. Let $\mathcal{D}_w$ denote the set of eligible training days belonging to weekday type w. Within each weekday type, we compute a weighted average wear probability across the 96 time slots, applying exponential recency weighting (half-life $H = 60$ days) so that more recent days contribute more to the baseline:

$$b_w(s) = \sum_{i \in \mathcal{D}_w} \tilde{\alpha}_i \, x_i(s), \qquad \alpha_i = \exp\left(-\frac{\ln 2}{H} \, a_i\right), \quad (1)$$

where $x_i(s) \in \{0, 1\}$ indicates wear at slot $s \in \{1, \ldots, 96\}$ on day i, $a_i = t^* - d_i$ is the age in days of training day d_i relative to the target day t^*, and $\tilde{\alpha}_i$ are the within-group normalized weights ($\sum_{i \in \mathcal{D}_w} \tilde{\alpha}_i = 1$). The resulting profile is smoothed with a Gaussian filter ($\sigma = 1.5$ slots, circular) to reduce noise, and represents the probability of observing device wear at given times of the day for each day-type group.

Anomaly Detection

Our pipeline (Figure 1) compares each test day's wear pattern against a personalized, weekday-matched baseline using Jensen-Shannon divergence.

For each day in the test window (60 days before to 30 days after the visit), we compare the observed wear pattern against the corresponding day-type baseline using the Jensen-Shannon divergence (JSD). Each 15-minute time slot is treated as a Bernoulli observation over wear vs. non-wear, and JSD contributions are summed across all 96 slots to yield a single per-day score:

$$\text{JSD}(o, b_w) = \tfrac{1}{2} \sum_{s=1}^{96} \Big[\text{KL}(o(s) \,\|\, m(s)) + \text{KL}(b_w(s) \,\|\, m(s)) \Big], \quad (2)$$

where $o(s)$ is the observed wear pattern of the test day, $m(s) = \frac{1}{2}(o(s) + b_w(s))$ is the pointwise mixture, and KL denotes the Bernoulli Kullback–Leibler divergence. A day is flagged as anomalous if its divergence score exceeds the 90th percentile of scores observed across the baseline training days for that day type:

$$\tau_w = \text{Percentile}_{90}(\{\text{JSD}(x_i, b_w) : i \in \mathcal{D}_w\}). \quad (3)$$

This threshold is personalized and adapts to each individual's natural variability in wear behavior.

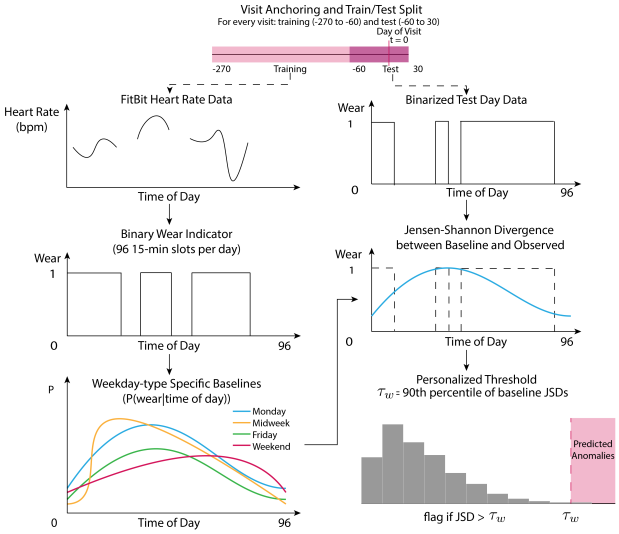

Figure 1. Overview of the personalized wear pattern anomaly detection pipeline. Fitbit heart rate is binarized into 15-minute wear indicators, compared against a weekday-matched baseline via Jensen-Shannon divergence, and flagged as anomalous when the divergence exceeds a personalized 90th-percentile threshold.

Evaluation

We adopt a visit-anchored evaluation, aggregating results across all included visits. For each day offset relative to the visit (from -60 to +30), we compute the proportion of visits with an anomaly on that day. We report anomaly rates for four windows: far baseline (days -60 to -30), near-visit (days -7 to 0), visit day (0), and post-visit (days +1 to +7).

Robustness. We vary the training window length (30, 90, 150, and 210 days) to evaluate sensitivity to the amount of historical data available for baseline estimation.

Results

Anomaly Probability Around Cardiovascular Visits

Figure 2 shows the primary result. The anomaly probability is stable at approximately 6.46% during the far baseline period (days -60 to -30), then rises beginning around day -10. In the week before the visit, the anomaly rate reaches 9.39% (1.45x the far baseline). On the visit day itself, the rate is 10.6%, and in the week after the visit, it is 12.9% (2.00x the far baseline). The 95% confidence interval separates from the baseline level around day -5, and the elevated rate persists through the full 30-day post-visit window.

Aggregate Missingness Does Not Capture the Signal

To contextualize the anomaly detection results, we compare aggregate missingness rates across the same periods (Figure 3). Mean daily missingness is 41.4% during the baseline period, 41.8% in the week before (Welch's t-test, p = 0.87, d = +0.01), and 45.0% in the week after (p = 0.072, d = +0.09). Neither the pre-visit nor the post-visit change in aggregate missingness reaches statistical significance, yet the anomaly detector identifies a clear signal over the same interval. This dissociation suggests that the behavioral change preceding visits involves shifts in when during the day wear occurs (e.g., altered sleep times, skipped mornings), rather than a simple reduction in total wear time.

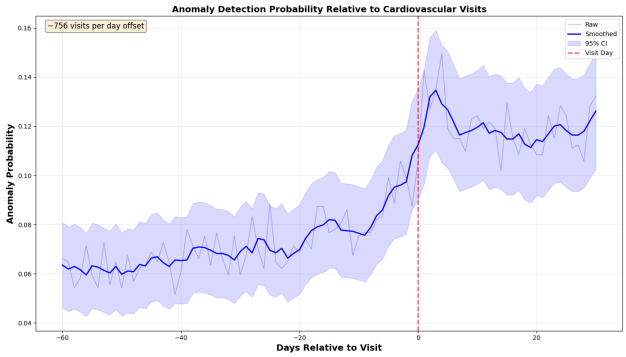

Figure 2. Predicted anomaly probability relative to cardiovascular emergency department visits (n = 756 visits). The shaded region shows the 95% confidence interval. Anomaly probability rises beginning approximately 10 days before the visit and remains elevated afterward.

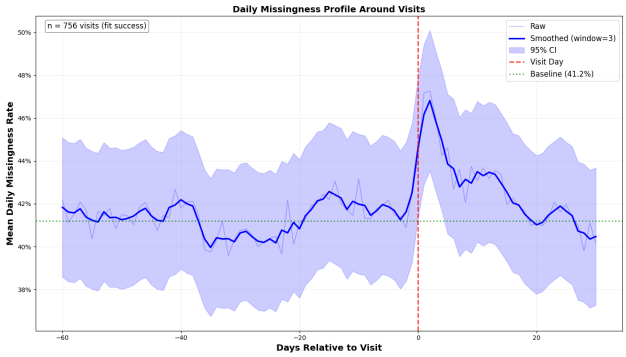

Figure 3. Missingness fraction alone relative to cardiovascular emergency department visits. Aggregate missingness shows no statistically significant change in the week before (p = 0.87) or week after (p = 0.072) the visit, supporting the necessity of incorporating diurnal dependence in anomaly detection.

Robustness to Training Window Length

Figure 4 shows anomaly probability curves under four training configurations. Because the training and test windows together determine the minimum required Fitbit enrollment, shorter training windows admit more visits: 998 visits qualify at 30 days of training (90-day total span) compared to 756 at 210 days (270-day total span). Thus, shorter configurations test on a progressively more inclusive cohort that includes participants with briefer enrollment periods. With only 30 days of training, the baseline is poorly estimated and

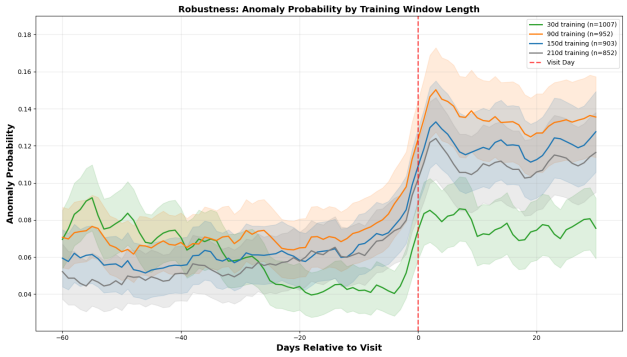

Figure 4. Sensitivity analysis for anomaly rate with varying training window lengths. While 30 days appears insufficient for distinguishing an anomaly around the cardiovascular event, the anomaly prediction is consistent across 90-210 days.

the detector fails to produce a meaningful signal (near/far ratio = 0.90x). At 90 days and above, the signal is consistently present, with near/far ratios of 1.30x, 1.29x, and 1.45x for 90, 150, and 210 days of training, respectively. The relationship between training length and detector behavior involves two competing effects. Longer training windows reduce the far-baseline anomaly rate (from 7.4% at 30 days to 6.5% at 210 days), reflecting better-calibrated baselines with fewer false positives. However, shorter training windows (90 days) yield the highest absolute anomaly rates in the near-visit and post-visit periods (11.9% and 16.5%, compared to 9.4% and 12.9% at 210 days). A shorter baseline captures more recent wear behavior, making the detector more reactive to acute deviations, at the cost of a noisier baseline. All configurations from 90 to 210 days produce a consistent temporal pattern: a stable baseline far from the visit, a rise beginning approximately 10 days before the visit, and sustained elevation afterward. The consistency of this pattern despite varying both the detector's training data and the included cohort supports the robustness of the finding.

Discussion

Individuals tend to use their wearable devices with repeatable habits at the day and week level, e.g. day-wear versus night-wear, and delayed mornings on weekends (Hershkovich et al., 2025). In this analysis, we demonstrate that this cyclical habituality property of wearable device usage can be leveraged to indicate when acute anomalies are occurring without utilizing any of the measurement values (steps, heart rate, etc.) from the device. Ignoring all signals except for wear/non-wear means that our algorithm can predict when data are missing not at random (MNAR), which is a traditionally difficult-to-impossible task in most data modalities (Rubin, 1976; Little, 2021; Ma & Zhang, 2021).

When an acute medical event occurs, an individual who should be indicating anomalous readings on their wearable device may stray from their consistent wearable device usage, leading to this signal not being measured (Hurwitz et al., 2026; Attig & Franke, 2020). For example, when a person contracts an illness like influenza or COVID, we may expect to measure raised body temperatures and lowered heart rate variability (Goergen et al., 2022; Grzesiak et al., 2021), but they may also not put on their smartwatch because they don't feel well, so those values would not be measured (MNAR). Approaches for data imputation could naively impute using past data outside of the infection, which could falsely imply in the data that everything is normal. Our algorithm can be used to indicate periods of likely anomalies in wear given an individual's historic habits, giving another dimension to inference. We utilized cardiovascular visits to an emergency department as an example of acute medical event; future work should validate these results with other acute events like infection.

Leveraging habituality, our algorithm shows a clear rise in anomalous wear behavior before, during, and after emergency visits for cardiovascular purposes (Figure 2). We demonstrate that missingness level alone (i.e. fraction of the day that the device was worn) is not sufficient to indicate that a prolonged anomaly has occurred (Figure 3), which validates that the time of day and day of the week are necessary context for examining missingness. Interestingly, while Figure 2 shows that many people change their wear habits around the cardiovascular event, the proportion does not rise to 100%—some people appear to continue to wear their device consistently throughout. Future research can consider using these individuals to guide imputation for those with predicted MNAR, for example through a pattern-mixture model (Little, 2025).

In future work, we aim to develop acute cardiovascular event warning algorithms that combine numerical vital-sign signals and wear patterns.

Acknowledgments

We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data [and/or samples and/or cohort] examined in this study.

References

Attig, C. and Franke, T. Abandonment of personal quantification: A review and empirical study investigating reasons for wearable activity tracking attrition. *Computers in Human Behavior*, 102:223–237, 2020.

Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G., Dishman, E., and The All of Us Research Program Investigators. The "All of Us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

Dhingra, L. S., Aminorroaya, A., Oikonomou, E. K., Nargesi, A. A., Wilson, F. P., Krumholz, H. M., and Khera, R. Use of wearable devices in individuals with or at risk for cardiovascular disease in the US, 2019 to 2020. *JAMA Network Open*, 6(6):e2316634, 2023.

Goergen, C. J., Tweardy, M. J., Steinhubl, S. R., Wegerich, S. W., Singh, K., Mieloszyk, R. J., and Dunn, J. Detection and monitoring of viral infections via wearable devices and biometric data. *Annual review of biomedical engineering*, 24(1):1–27, 2022.

Grzesiak, E., Bent, B., McClain, M. T., Woods, C. W., Tsalik, E. L., Nicholson, B. P., Veldman, T., Burke, T. W., Gardener, Z., Bergstrom, E., et al. Assessment of the feasibility of using noninvasive wearable biometric monitoring sensors to detect influenza and the common cold before symptom onset. *JAMA network open*, 4(9):e2128534, 2021.

Hershkovich, L., Jeong, H., Sakai, S., Kane, H., Allen, A., Goldstein, B., and Dunn, J. Defining the habitome: Phenotypes of routine and their relationship to health outcomes. Research Square preprint, doi: 10.21203/rs.3.rs-5861743/v2, 2025.

Hurwitz, E., Connelly, E., Sklerov, M., Master, H., Hochheiser, H., Butzin-Dozier, Z., Dunn, J., and Haendel, M. A. Population differences in wearable device wear time: Rescuing data to address biases and advance health equity. medRxiv preprint, doi: 10.64898/2026.03.06.26347799, 2026.

Lederer, L., Breton, A., Jeong, H., Master, H., Roghanizad, A. R., and Dunn, J. The importance of data quality control in using Fitbit device data from the All of Us research program. *JMIR mHealth and uHealth*, 11:e45103, 2023.

Little, R. J. Missing data assumptions. *Annual Review of Statistics and Its Application*, 8(1):89–107, 2021.

Little, R. J. Pattern-mixture models for missing data. *JAMA*, 334(23):2126–2127, 2025.

Ma, C. and Zhang, C. Identifiable generative models for missing not at random data imputation. *Advances in Neural Information Processing Systems*, 34:27645–27658, 2021.

Nijman, S. W. J., Leeuwenberg, A. M., Beekers, I., Verkouter, I., Jacobs, J. J. L., Bots, M. L., Asselbergs, F. W., Moons, K. G. M., and Debray, T. P. A. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of Clinical Epidemiology*, 142:218–229, 2022.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976.