# DETECTING PERIODIC BIASES IN WEARABLE-BASED ILLNESS DETECTION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Wearable health devices have revolutionized our ability to continuously analyze human behavior and build longitudinal statistical models around illness by measuring physiological indicators like heart rate over several months of an individual's life. Shifts in these indicators have been correlated with the onset of illnesses such as COVID-19, leading to the development of Wearable-Based Illness Detection (W-BID) models that aim to detect the onset of illness. While W-BID models accurately detect illness, they often over-predict illness during healthy time periods due to variance caused by seemingly random human choices. However, it is because W-BID models treat each input window as independent and identically distributed samples that we are unable to account for the weekly structure of variance that causes false positives. Towards preventing this, we propose a system for identifying structural variance in wearable signals and measuring the effect they have on W-BID models. We demonstrate how a simple statistical model that does not account for weekly structure is strongly biased by weekly structure, with a Pearson correlation coefficient of 0.9.

## 1 INTRODUCTION

Wearable health devices such as the Apple Watch, Fitbit and Garmin Watch have started a movement towards personalized and predictive health tools Purawat et al. (2021); Li et al. (2017). While these devices currently provide suggestions around light exercise, research has shown the potential for the detection and prediction of illnesses such as COVID-19 Dunn et al. (2021); Radin et al. (2020); Kavsaoğlu et al. (2015); Mason et al. (2022). These statistical models have the potential to revolutionize how the general public makes decisions around when and why to initiate visits to medical professionals.

Key to the empowerment of these wearable device users is establishing trust in the reliability of our statistical methods. One of the biggest threats to users' and medical professionals' trust is high false positive rates. While prior work has shown the potential for these statistical methods, many models show false positives before the sickness period for the individual and struggle to perform on certain individuals Abir et al. (2022); Merrill & Althoff (2022). When we consider the vast landscape of differences in human physiology, it is natural that there are individuals for whom it is more difficult to detect illness. The variance of different peoples' lifestyles and bodies is a key deterrent in the wide-scale application of Wearable-Based Illness Detection models.

A likely explanation for pre-illness false positives is the similarity in the presentation of common weekend activities, such as sleeping late and drinking alcohol, and illness. For example, COVID-19 can be seen as an increase in heart rate prior to symptom onset Natarajan et al. (2020). However, poor sleep and consumption behaviors common to weekends also cause increased heart rate. As a result, we hypothesize that W-BID models, which are mostly trained with the assumption that each sample is independent and identically distributed, are periodically affected by the increased heart rate that could also signal illness during some regular portion of the week.

To press this issue, we propose exploiting the time-domain structure of human behaviors towards handling this variance, specifically the 7-day weekly structure that we organize our decisions around. While humans make different decisions every day, we often make decisions around exercise, consumption, and sleep based on the day of the week. Weekends are often times for increased sleep, and unstructured schedules, while weekdays are more often associated with structured work and eating

routine To et al. (2022); Burchartz et al. (2022); Esposito et al. (2022). Although the specific patterns vary depending on an individual's particular life conditions, this intuition that individuals have weekly "rhythms," we believe, is a strong prior for a majority of the population. If this structure has an effect on a person's physiological measures, then we believe it affects W-BID model predictions during variant times of the week.

In this work, we aim to illustrate the degree to which human time series structures manifest in statistical models that aim to detect illness from wearable devices. To do this we will (1) illustrate the degree to which individuals have structural variance in their time series and (2) show how a simple statistical model is affected by these. To analyze the structure of weekly rhythms, we characterize how an individual's physiological indicators present differently on weekends versus weekdays, and the degree to which a strong effect in one's physiological indicator can correlate with a similar rhythm in the probability of illness predicted by a statistical model.

## 2 METHODS

### 2.1 DATA COLLECTION AND PREPROCESSING

For our analysis, we use the publicly available dataset described and used by Alavi et al.Alavi et al. (2022) which is available to download at this link: `https://storage.googleapis.com/gbsc-gcp-project-ipop_public/COVID-19-Phase2/COVID-19-Phase2-Wearables.zip`. We filter the time series from midnight to 7 AM each night as done in Alavi et al. Of the 73 participants who are infected by COVID-19 in this dataset, we aim to only analyze those who consistently have data for every day of the week. Towards this, we select the 29 individuals who are missing less than 5% of the time series data between midnight and 7 am.

### 2.2 WEEKLY STRUCTURE EXTRACTION

As part of the anonymization of this dataset, the COVID-19 diagnosis date is changed. As a result, we cannot rely on the day of the week being accurate, so, we look at the relative differences between days of the week to extract the weekly structure. To analyze the strength of weekly structures, we take advantage of the actogram visualization techniques from chronobiologyOike et al. (2019). We organize our time series into tiles of 10,080 minutes ($7\frac{\text{days}}{\text{week}} \times 24\frac{\text{hours}}{\text{day}} \times 60\frac{\text{minutes}}{\text{hour}}$). This represents 7 days of heart rate data where there is an average heart rate value for each minute. We stack 7-day tiles on top of each other in a heat map visualization to illustrate patterns along the same day of the week. We refer to this organization and visualization as a weekly actogram.

To analyze the strength of the weekly structure we see in our weekly actograms, we use the measure for effect size Cohen's D

$$\text{Cohen's } d = \frac{\text{mean(group A)} - \text{mean(group B)}}{\sqrt{\text{var(group A)} + \text{var(group B)}}}$$

Cohen's d indicates how many standard deviations one group's mean is from the other's. Because we know that most people organize their weeks around weekdays and weekends, we loosely quantify weekly structure as the effect size between the weekend and weekday values for an individual. These values can be the minute-level heart rate averages or the hour-level statistical model outputs. Cohen's d is often divided into small, medium, and large effects based on the thresholds 0.2, 0.5, and 0.8. To understand how significant this effect is, we perform the Mann-Whitney U test for the difference in distributions between the nightly means of the weekday and weekend values.

### 2.3 STATISTICAL CLASSIFIER TRAINING AND EVALUATION

Participants reported the date on which they took a positive COVID-19 test (DX date). The datasets were set up to include 7-day windows, starting at midnight of each day before the COVID-19 test date. 7-day windows were labeled as negative if they ended at least 5 days before the DX date. The windows were labeled as positive if they ended on the night before, or the night of the DX date, since the positive test indicated the participants were probably sick beforehand. For the same reasoning as above, the windows ending between 5 days before to 2 days before the DX date were not included

in either the positive or negative categories. One participant did not have enough baseline region to be included in the analysis.

We train a 1-layer Fully Connected Neural Network to perform classification on these 7-day windows, baselined using the period 65 to 35 days prior to illness onset, using the Pytorch library. While this model output binary predictions, we use the illness logit from the final layer. These logits are unbounded and represent how strongly the model would have predicted the output class. To understand the degree to which weekly rhythms in one's physiological data affect model performance, we organize the model outputs or logits as weekly actograms. We analyze the strength of the weekly structure in the model logits using Cohen's d. We evaluate a validation set of 9 participants who were randomly selected after we blocked by effect size. For our final analysis, we include all individuals.

We evaluate the model bias towards weekly structure by plotting the weekly effects in heart rate against the weekly effects in model logits and estimating the Pearson correlation coefficient. A Pearson correlation coefficient between 0.8-1.0 would indicate that the model has a strong bias due to weekly structure whereas a coefficient between 0.0-0.4 would indicate a weak or no bias.

# 3 RESULTS

## 3.1 PHYSIOLOGICAL DATA

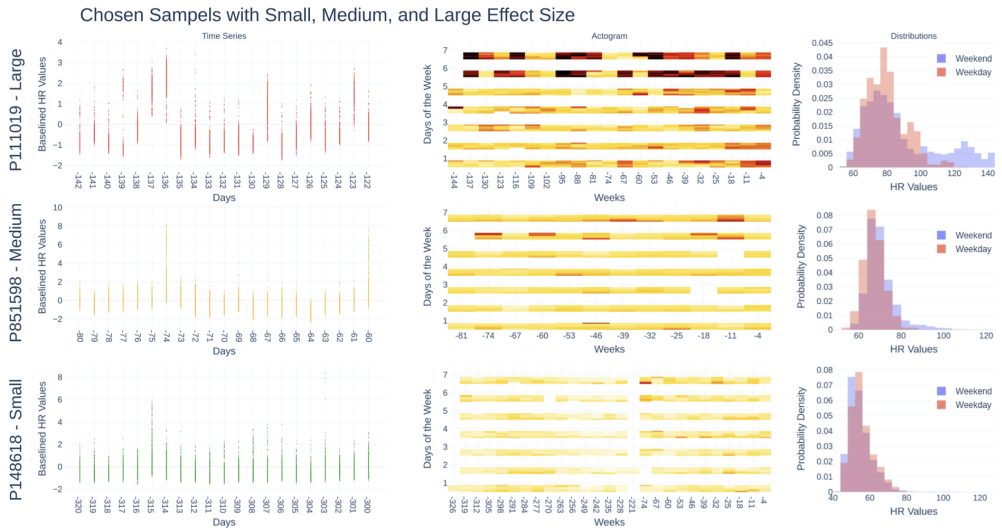


Figure 1: Chosen participants from the three effect size groups. Left: Example Time series Center: Example weekly actograms Right: Weekday vs. weekend distributions

We illustrated the strength of weekly structures in Figure 1. The top participant shows strong effects with clear spikes in their time series, a darkened top two rows of their actogram indicating elevated heart rates, and a large right shift in the weekend histogram relative to the weekday. The middle participant shows this to a degree, while the bottom participant illustrates having no weekly structure. In Figure 2, we show all participants' Cohen's d and p-values for the minute-level heart rate. 18 out of 29 participants showed small to large weekly effects.

## 3.2 ILLNESS PROBABILITIES

Here we show the weekly structure in the model logits. The scatter plot of effect sizes for heart rate versus model logits illustrates the correlation between these effect. The Pearson Correlation Coefficient of 0.9 indicating that the model's predictions are strongly biased by the weekly structure.

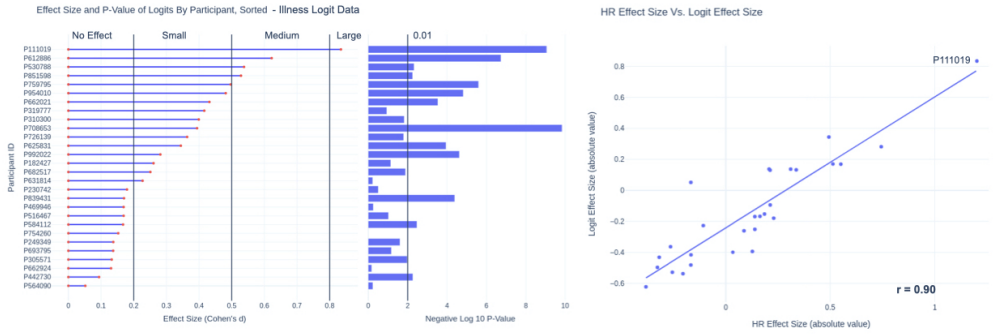


Figure 2: Left: All participant's weekday vs weekend illness logit effect sizes sorted from strongest weekend over weekday effects to strongest weekday over weekend effects. Center: Negative log p-value of the effect for each participant. Higher values represent more significant effects by orders of magnitude. Right: Scatter plot of the Heart Rate effect size versus Illness Logit effect size. Pearson correlation coefficient of 0.9.

## 4 Discussion

In this work, we demonstrated how weekly structures can emerge in wearable health data and how these structures can bias the predictions of statistical models. We present a system for identifying when weekly structures emerge so that they can be accounted for in illness detection systems. This system can be used to assess the degree to which a W-BID model is biased by individual weekly structure. Our results suggest that individuals have their own weekly structures that meaningfully shift W-BID model predictions and that many systems may unintentionally be biased by these weekly structures. Future systems should track structure in time series at different time scales and use them to adjust predictions.

## References

Farhan Fuad Abir, Khalid Alyafei, Muhammad EH Chowdhury, Amith Khandakar, Rashid Ahmed, Muhammad Maqsud Hossain, Sakib Mahmud, Ashiqur Rahman, Tareq O Abbas, Susu M Zughaier, et al. Pcovnet: A presymptomatic covid-19 detection framework using deep learning model using wearables data. *Computers in biology and medicine*, 147:105682, 2022.

Arash Alavi, Gireesh K Bogu, Meng Wang, Ekanath Srihari Rangan, Andrew W Brooks, Qiwen Wang, Emily Higgs, Alessandra Celli, Tejaswini Mishra, Ahmed A Metwally, et al. Real-time alerting system for covid-19 and other stress events using wearable data. *Nature medicine*, 28(1): 175–184, 2022.

Alexander Burchartz, Doris Oriwol, Simon Kolb, Steffen CE Schmidt, Birte von Haaren-Mack, Claudia Niessner, and Alexander Woll. Impact of weekdays versus weekend days on accelerometer measured physical behavior among children and adolescents: results from the momo study. *German Journal of Exercise and Sport Research*, 52(2):218–227, 2022.

Jessilyn Dunn, Lukasz Kidzinski, Ryan Runge, Daniel Witt, Jennifer L Hicks, Sophia Miryam Schüssler-Fiorenza Rose, Xiao Li, Amir Bahmani, Scott L Delp, Trevor Hastie, et al. Wearable sensors enable personalized predictions of clinical laboratory measurements. *Nature medicine*, 27(6):1105–1112, 2021.

Francesco Esposito, Francesco Sanmarchi, Sofia Marini, Alice Masini, Susan Scrimaglia, Emanuele Adorno, Giorgia Soldà, Fabrizio Arrichiello, Filippo Ferretti, Marilisa Rangone, et al. Weekday and weekend differences in eating habits, physical activity and screen time behavior among a sample of primary school children: the "seven days for my health" project. *International journal of environmental research and public health*, 19(7):4215, 2022.

A Reşit Kavsaoğlu, Kemal Polat, and Muthusamy Hariharan. Non-invasive prediction of hemoglobin level using machine learning techniques with the ppg signal's characteristics features. *Applied Soft Computing*, 37:983–991, 2015.

Xiao Li, Jessilyn Dunn, Denis Salins, Gao Zhou, Wenyu Zhou, Sophia Miryam Schüssler-Fiorenza Rose, Dalia Perelman, Elizabeth Colbert, Ryan Runge, Shannon Rego, et al. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLoS biology*, 15(1):e2001402, 2017.

Ashley E Mason, Frederick M Hecht, Shakti K Davis, Joseph L Natale, Wendy Hartogensis, Natalie Damaso, Kajal T Claypool, Stephan Dilchert, Subhasis Dasgupta, Shweta Purawat, et al. Detection of covid-19 using multimodal data from a wearable device: results from the first tempredict study. *Scientific reports*, 12(1):3463, 2022.

Mike A Merrill and Tim Althoff. Self-supervised pretraining and transfer learning enable flu and covid-19 predictions in small mobile sensing datasets. *arXiv preprint arXiv:2205.13607*, 2022.

Aravind Natarajan, Hao-Wei Su, and Conor Heneghan. Assessment of physiological signs associated with covid-19 measured using wearable devices. *NPJ digital medicine*, 3(1):156, 2020.

Hideaki Oike, Yukino Ogawa, and Katsutaka Oishi. Simple and quick visualization of periodical data using microsoft excel. *Methods and protocols*, 2(4):81, 2019.

Shweta Purawat, Subhasis Dasgupta, Jining Song, Shakti Davis, Kajal T Claypool, Sandeep Chandra, Ashley Mason, Varun Viswanath, Amit Klein, Patrick Kasl, et al. Tempredict: a big data analytical platform for scalable exploration and monitoring of personalized multimodal data for covid-19. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 4411–4420. IEEE, 2021.

Jennifer M Radin, Nathan E Wineinger, Eric J Topol, and Steven R Steinhubl. 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, 2020.

Quyen G To, Robert Stanton, Stephanie Schoeppe, Thomas Doering, and Corneel Vandelanotte. Differences in physical activity between weekdays and weekend days among us children and adults: Cross-sectional analysis of nhanes 2011–2014 data. *Preventive Medicine Reports*, 28: 101892, 2022.