UNIVERSITY OF CALIFORNIA SAN DIEGO

# Towards the Use of Commercial Wearable Devices for Acute Infectious Disease Mitigation

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioengineering

by

Patrick Kasl

Committee in charge:

Professor Benjamin Smarr, Chair
Professor Cinnamon Bloss
Professor Gert Cauwenberghs
Professor Berk Ustun

2024

The Dissertation of Patrick Kasl is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

# EPIGRAPH

*Life is so rich.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to acknowledge my advisor, Benjamin Smarr, for his never-ending support and belief in me. His patience, guidance, and humor throughout this process has kept me sane. I certainly would not be here had he not taken a chance on me when we first started working together. I would also like to thank my colleagues and friends in the Smarr Lab. When I needed a sounding board for my antics, they were there. When I needed feedback on a submission, they were there. While I hope to be able to repay them, I say this with the knowledge that I never truly will be able to so.

I would also like to thank everyone in San Diego for the time we spent together, whether it was in line at Kikos, picking up absurdly heavy objects only to put them back down, or the various discussions on any number of our meandering walks. Thank you for letting me be a part of this chapter in your story.

Chapter 2, contains information as it appears in A.E. Mason, P. Kasl, W. Hartogensis, J.L. Natale, S. Dilchert, S. Dasgupta, S. Purawat, A. Chowdhary, C. Anglo, D. Veasna, L.S. Pandya, L.M. Fox, K.Y. Puldon, J.G. Prather, A. Gupta, I. Altintas, B.L. Smarr, and F.M. Hecht, "Metrics from Wearable Devices as Candidate Predictors of Antibody Response Following Vaccination against COVID-19: Data from the Second TemPredict Study", 2022 *Vaccines*. The dissertation author was a co-first author on this paper. It also contains information as it appears in P. Kasl, L. Keeler Bruce, W. Hartogensis, S. Dasgupta, L.S. Pandya, S. Dilchert, F.M. Hecht, A. Gupta, I. Altintas, A.E. Mason, and B.L. Smarr, "Utilizing wearable device data for syndromic surveillance: A fever detection approach", 2024 *Sensors*. The dissertation author was the primary author of this paper.

Chapter 3, contains information as it appears in P. Kasl, L. Keeler Bruce, W. Hartogensis, S. Dasgupta, L.S. Pandya, S. Dilchert, F.M. Hecht, A. Gupta, I. Altintas, A.E. Mason, and B.L. Smarr, "Utilizing wearable device data for syndromic surveillance: A fever detection approach", 2024 *Sensors*. The dissertation author was the primary author of this paper.

Chapter 4, is a reprint of P. Kasl, S. Soltani, L. Keeler Bruce, V. Kumar Viswanath, W.

Hartogensis, A. Gupta, I. Altintas, S. Dilchert, F.M. Hecht, A.E. Mason, and B. L. Smarr, "A Cross-study Analysis of Wearable Datasets and the Generalizability of Acute Illness Monitoring Models", 2024 currently in press at the *Conference on Health Inference and Learning*. The dissertation author was the primary author of this paper.

## VITA

2020 Bachelor of Science, University of Wisconsin–Madison

2024 Doctor of Philosophy, University of California San Diego

## PUBLICATIONS

**P. Kasl**, S. Soltani, L. K. Bruce, et al., "A Cross-study Analysis of Wearable Datasets and the Generalizability of Acute Illness Monitoring Models," in press at the *Conference on Health, Inference, and Learning*, 2024

**P. Kasl**, L. K. Bruce, W. Hartogensis, et al., "Utilizing wearable device data for syndromic surveillance: A fever detection approach," *Sensors*, 2024

A. E. Mason*, **P. Kasl***, W. Hartogensis, et al., "Metrics from Wearable Devices as Candidate Predictors of Antibody Response Following Vaccination against COVID-19: Data from the Second TemPredict Study," *Vaccines* 2022

A. E. Mason, **P. Kasl**, S. Soltani, et al., "Elevated body temperature is associated with depressive symptoms: Results from the TemPredict Study," *Scientific Reports* 2024

L. K. Bruce, **P. Kasl**, S. Soltani, et al., "Variability of temperature measurements recorded by a wearable device by biological sex," *Biology of Sex Differences*, 2023

S. Purawat, S. Dasgupta, J. Song, et al., "TemPredict: A Big Data Analytical Platform for Scalable Exploration and Monitoring of Personalized Multimodal Data for COVID-19," *2021 IEEE International Conference on Big Data (Big Data)*, 2021

H. Kletzien, S. M. Wang, **P. Kasl**, and N. P. Connor, "Lingual Muscle Plasticity with Age and Exercise.," *Dysphagia* 2022

ABSTRACT OF THE DISSERTATION

Towards the Use of Commercial Wearable Devices for Acute Infectious Disease Mitigation

by

Patrick Kasl

Doctor of Philosophy in Bioengineering

University of California San Diego, 2024

Professor Benjamin Smarr, Chair

The landscape of health technologies is rapidly evolving and commercial wearable devices equipped with health sensors offer a promising avenue for continuous health monitoring. The ubiquity of these devices among consumers presents an unprecedented opportunity to leverage the wealth of data that wearables collect for health applications. Despite their widespread use, there have only recently been developments towards utilizing these data for enhancing health monitoring. However, research stemming from recent efforts to gather large-scale longitudinal wearable datasets suggests these devices might hold potential in detecting the presence of and characterizing aspects of acute physiological changes. The application of wearable device data might be particularly useful in the context of acute physiological changes given wearables'

ability to monitor a number of physiological vital signs both passively and longitudinally. Passive monitoring uniquely enables longitudinal comparisons of individuals to themselves over time and real-time identification of significant deviations indicative of changes in health status. This thesis explores the application of data from wearable devices for detecting and characterizing physiological changes following significant health events, specifically vaccination for COVID-19 and the onset of fever. I also present the first comprehensive analysis of multiple large-scale, longitudinal wearable device datasets, therein assessing the generalizability of algorithms for monitoring acute illnesses and characterizing the biases in these datasets, some of which are correlated with demographic variables. Through this research, I demonstrate the potential of commercial wearable devices in enhancing our understanding and monitoring of acute physiological changes, and present a framework through which industry and research standards might emerge to speed the evolution of this field.

# Chapter 1

# Introduction

Quantification is central to scientific inquiry; broadly, empirical progress is driven by better measurements (Houle et al., 2011). Measurement capacities often either improve through advancements in theory, pointing towards quantities of relative import, or better instruments, improving measurements of familiar quantities or facilitating those previously unknown (Weimer, 2023). The work in this thesis deals primarily with the latter and was enabled by the adoption of mass-produced wearable devices which quantify human physiology at unprecedented scales. Such large-scale quantification efforts have been a cornerstone in advancing our understanding of biological systems. Where, for example, genomic sequencing has opened new routes to precision medicine, data from wearables might enable new paradigms for health management. Ultimately, the inherent complexity of biological systems destines models developed using biological data to be "impoverished" relative to the phenomena they describe (Weimer, 2023). That is, measurements on any particular in vivo biological system provide information about one specific instantiation of many possible ones (e.g., organisms vary genetically and phenotypically depending on that organism's unique history) and any set of such measurements are nearly always correlated. Unresolvable "latent" or "hidden" variables often drive these correlations such that any model, either theoretical or learned in some data-driven way, might never fully capture the inherent complexity of the system they intend to describe. Humans, in particular, exist within this context of dynamically changing constraints. However, leveraging time-dependent

measurements, like those made by wearable devices, might provide a means to meaningfully reduce some of this complexity.

Wearable devices are an expanding category of devices that are increasingly adopted by consumers (Huhn et al., 2022). While initially favored by individuals with a keen interest in characterizing their own physiology, the proliferation of affordable, multifunctional wearables has resulted in millions of individuals with devices capable of monitoring several physiological vital signs simultaneously, longitudinally, and with high levels of accuracy. Despite the vast data these devices collect, their potential for advanced health-tracking applications remains largely untapped. This is evolving as both researchers and device manufacturers have gained access to large-scale, longitudinal wearable device datasets and demonstrated the utility of wearable devices for physiological monitoring with applications in both within-individual and public health-level monitoring (Radin et al., 2020). Recent works have demonstrated feasible implementations in mental health (Xu et al., 2022a; Wainberg et al., 2021), acute physiology (Goergen et al., 2022; Abir et al., 2022; Richards et al., 2021; Gadaleta et al., 2021; Conroy et al., 2022; Yamagami et al., 2021; Hirten et al., 2021; Natarajan et al., 2020; Mayer et al., 2022; Alavi et al., 2022; Miller et al., 2020; Pho et al., 2023; Hirten et al., 2022; Merrill et al., 2023; Grzesiak et al., 2021; Mezlini et al., 2022; Chaudhury et al., 2022), and beyond (Lam et al., 2021; Master et al., 2022). However, the transition from research findings to practical applications presents notable challenges, as with many other emerging biotechnologies (Drolet and Lorenzi, 2011). Regardless of the immediate outcomes of ongoing research efforts, the utility these devices present to consumers likely ensures their continued use. As such, current and future research efforts will be crucial for demonstrating how data from these devices can have broader health impacts beyond those that benefit any individual user.

Throughout this work, I present novel characterizations of acute physiological changes as measured by wearable devices. I further show how these changes can be systematically leveraged to build models capable of detecting those physiological changes at scale. Ultimately, I hope this work demonstrates the utility of this relatively new data type to aid in optimizing individual and

organizational-level responses to persistent acute physiological challenges, despite substantial barriers to real-world implementations of such systems. In the remainder of this introduction, I provide a brief overview of the working principle behind the wearables used in this research, some basic descriptions of the human physiological mechanisms underpinning the phenomena I characterize, and the statistical and machine-learning techniques used throughout the thesis.

## 1.1 Wearable devices

Throughout this work, I consider a class of devices (i.e., so-called consumer, off-the-shelf, commercially available, etc.) available to consumers on a mass-market basis. Such devices are often referred to by the terms "fitness trackers", "activity trackers", and "smartwatches" which, for the sake of this work, I take to be synonymous. The key distinction between the class of devices I consider and other consumer devices is the device's ability to measure some aspect of human physiology across time. I use the term *wearable* to refer to such devices for the sake of this work, however, this category of devices is broad and continuously growing. In particular, in this work I consider the subset wearables with the capacity to quantify some notion of activity, often by leveraging data from accelerometers, and heart rate, often leveraging data from photoplethysmography (PPG) sensors. Contemporary examples include the Apple Watch and devices manufactured by FitBit which are based on the watch form factor. Much of my work also focuses on data from the Oura Ring (Oura Ring, Oura Health Oy, Oulu, Finland), a wearable based on the finger-ring form factor. The development of novel form factors and the inclusion of novel sensors in wearables is an important area of research. Progress in this domain might enable an improved understanding of human physiology and more accurate models based on wearable data. Arguably, the key factor that makes data from these devices an interesting topic of research is their ability to capture human physiology unobtrusively, longitudinally, and at unprecedented scale. *Longitudinal* within-individual data capture enables comparisons of individuals to themselves over time, which makes it possible to detect physiological changes

that might have previously been dismissed as noise when measured across a large population. Similarly, gathering *large-scale* longitudinal physiological data allows the characterization of innate physiological phenomena which regress to an interpretable mean above some noise floor not readily apparent at smaller sample sizes, as I will describe in Chapters 2 and 3.

Wearable devices employ a suite of sensors to measure human physiology across time. The wearables used in this work process data from these sensors using proprietary algorithms to extract certain physiological metrics. An important distinction here is that because these devices are aimed at a consumer-facing market (as opposed to, for example, a strictly research-based community), these extracted metrics are informed by choices that make those metrics useful to the average consumer. These choices exist within the context of technological tradeoffs optimized to sell a product to consumers, and as such, are not necessarily developed with the end goal of providing research-grade information. Accordingly, device designers and engineers might seek a minimax over price and features. Of these features, physiological monitoring is merely a subspace and might need to compete, in a technical development, physical space, and energy consumption sense, against other features important to consumers (e.g., text messaging). Any of the metrics reported by any particular wearable may well be an imperfect measure of an individual's underlying physiology and certain design choices (i.e., quantifying "steps per unit time") lead to a lossy compression of an individual's underlying physiology. Similarly, the wearables used in this work compress photoplethysmography data into an average heart rate, heart rate variability, and respiratory rate over a certain unit of time. Nonetheless, these devices measure these physiological vital signs with a high degree of accuracy (Shcherbina et al., 2017). While it's clear that data from these devices present useful information about an individual's physiology, as I will discuss in Chapter 4, unknown differences in each device's proprietary algorithms and hardware might present substantial barriers to a unified approach to utilizing these data. Indeed, much of Chapter 4 aims to be an initial step towards developing a framework for considering data from different devices in unison.

### 1.1.1 Accelerometers

The accelerometers used in commercial wearables are commonly based on microelectromechanical systems (MEMS) (Mohd-Yasin et al., 2003). These sensors typically consist of small mechanical systems integrated on a chip. Many consumer-grade accelerometers are based on capacitive sensing, where small, movable proof masses are coupled with springs (Mohd-Yasin et al., 2003). When an external force is applied to the sensor, the capacitance in the microstructure changes proportional to the acceleration applied to the sensor. Changes in capacitance over time allow quantification of the acceleration experienced by the wearable device. Many consumer-grade accelerometers quantify acceleration in three orthogonal axes (i.e., 3D accelerometry).

Data from accelerometers tends to exhibit substantial levels of noise and raw 3D accelerometry can be challenging to interpret directly. As such, most wearables pass raw accelerometry data through several filters to reduce noise and extract meaningful features from these data, such as the number of steps (Lee et al., 2017). Such methods commonly include low-pass filtering to remove high-frequency vibrations unrelated to human motion and various smoothing approaches which tend to reduce variability due to noise. After noise reduction, it is common to apply a threshold-based algorithm to detect signatures of "steps", however, several other approaches might be used to detect steps.

### 1.1.2 Photoplesmography sensors

While accelerometers are commonly included in a broad range of applications ranging from shipping labels to automotive and aeronautical sensors, photoplethysmography (PPG) sensors are more specific to wearable health applications. PPG technology can be configured in a wide variety of ways, however, principally, they are composed of a set of one or more light sources and a set of one or more photosensitive elements. Light emitted is typically in the range of 500 to 1000 nanometers (nm), however, in theory, any wavelength of light that

varies in concordance with the amount of blood in the tissue can be used in PPG sensors. In practice, many consumer-grade devices use green light ( 500 nm) with light-emitting diodes (LEDs) as a light source and a photodiode as the photosensitive element (Maeda et al., 2011). Photoresistors can also be used, however, for the sake of brevity I describe the more commonly used photodiode-based systems. PPG technology leverages the absorbance of light in this range by blood. This absorbance varies with time along with the pulse. After a heartbeat, blood travels through the vasculature to the periphery, where a PPG sensor is typically located (i.e., on the wrist in the case of watch form factors or finger in the case of ring form factors). The amount of light that hits the photodiode is correlated with the amount of blood in the tissue at that time, and the voltage produced by the photodiode over time is commonly referred to as a PPG waveform (OB1, 2022). In general, the sampling rate of commercial-grade PPG systems is on the order of 100 Hz, substantially higher than the Nyquist rate for the phenomena they intend to measure (i.e., an individual's average resting heart rate is typically on the order of 60 bpm or 1 Hz).

Peaks in the PPG waveform can be isolated using traditional digital signal processing approaches, typically following a set of low-pass filters to remove noise artifacts and the subsequent application of peak-finding algorithms (OB1, 2022). Again, the precise steps involved amount to a proprietary algorithm that is optimized for the configuration of any particular wearable device. Here, as an example, I provide the publicly available details for the Oura Ring, a wearable whose data I use extensively throughout this thesis.

The Oura Ring calculates heart rate (HR), heart rate variability (HRV), and respiratory rate (RR) from inter-beat intervals (IBIs) during periods of sleep. IBIs are calculated using raw PPG data processed using a real-time moving average filter (Altini and Kinnunen, 2021; Kinnunen et al., 2020). The local maximum and minimum values in these PPG data correspond to each heartbeat. The Oura Ring also estimates the probability that each IBI is an artifact. The Oura Ring uses a median filter to classify each IBI as either normal or abnormal; any individual IBI that is more than 16 beats per minute (bpm) removed from the median IBI in a moving window of length seven are marked as abnormal (Altini and Kinnunen, 2021; Kinnunen et al.,

2020). If any of the two IBIs before or after a particular IBI are abnormal, that set of five IBIs is not included in subsequent analyses. HR and HRV are calculated if at least 30% of the IBIs in a 5 min window are normal according to these criteria (Altini and Kinnunen, 2021; Kinnunen et al., 2020). The Oura Ring calculates HR using the mean IBI and HRV as the root mean square of successive differences (rMSSD). RR is calculated by finding peaks in IBI over the time period under analysis (Altini and Kinnunen, 2021; Kinnunen et al., 2020). These metrics are generated on-device and stored via the application while the raw PPG is not continuously recorded or stored for analysis.

### 1.1.3 Temperature sensors

Temperature sensors are increasingly being included in commercial wearables. At the initiation of the work in this thesis, temperature sensors were less common in these devices. However, the Oura Ring was one of the first to include one of these sensors. The temperature sensor in the Oura Ring consists of a thermistor, an electrical component whose resistance varies as a function of temperature. Temperature exhibits strong circadian rhythms; core temperature drops in the evening around the initiation of sleep while skin temperature rises around the same time, thus exhibiting structured variance at the same frequency but in inverse phases (Krauchi and Wirz-Justice, 1994). The Oura Ring measures temperature with two negative temperature coefficient (NTC) thermistors (non-calibrated, resolution of 0.07 degrees Celsius) located palmar when the ring is worn as intended and temperature readings are recorded at 1-minute intervals.

### 1.1.4 Sleep stages

Sleep staging has traditionally been approached using polysomnography (PSG) however recent developments have enabled the use of machine-learning to predict the stage of sleep an individual is in using sensor information available from wearable devices. As with accelerometry and PPG, proprietary algorithms are used to estimate the stage of sleep a user is in. Here, I describe the approach taken by the Oura Ring as an example. The Oura Ring calculates sleep

stages using a machine-learning classifier and predicts sleep stages on 30 second (s) windows of data (Altini and Kinnunen, 2021). The Oura Ring assesses temperature at 10 s intervals and samples less than 31 or more than 40 degrees Celsius are masked (Altini and Kinnunen, 2021). The mean, min, max, and standard deviation are calculated on a rolling basis (Altini and Kinnunen, 2021). High frequency accelerometry data are used to calculate the mean amplitude deviation (MAD) in 5 s windows (Altini and Kinnunen, 2021). The MAD represents the average deviation from the mean vector magnitude (Altini and Kinnunen, 2021). Within each 30 s window, the mean, max, and interquartile range (IQR) of MADs in the 10-90th percentile of the window are calculated (Altini and Kinnunen, 2021). The difference in arm angle was also calculated in each 5 s window and the mean, max, and IQR of arm angles in the 10-90th percentile of the 30 s window are calculated (Altini and Kinnunen, 2021). Processed accelerometry features are also calculated for each three individual axes (Altini and Kinnunen, 2021). High-resolution data are processed using a 5th-order Butterworth bandpass filter (3 to 11 Hz) and taking their absolute value (Altini and Kinnunen, 2021). The mean, max, and IQR of values in the 10-90th percentile within each axis are calculated for each 30 s window (Altini and Kinnunen, 2021). High-quality IBIs are identified in the same way they are identified for the calculation of HR and HRV in 5 min windows (Altini and Kinnunen, 2021). For each 30 s window, Oura calculates HR, HRV (rMSSD), and RR (Altini and Kinnunen, 2021). They also calculate the following additional HRV metrics: SDNN, pNN50, frequency power in the LF and HF bands, the main frequency peak in the LF and HF bands, total power, normalized power, mean and coefficient of variation in the zero-crossing interval. Excluding accelerometer-based features, each feature is normalized on a per-night basis using the 5-95 percentiles of that feature (Altini and Kinnunen, 2021). Oura claims this accounts for inter-individual differences in features.

In the preceding sections, I describe the processes through which the Oura Ring converts untransformed sensor data into specific features. These descriptions aim to highlight the specificity with which device manufacturers customize algorithms to their unique hardware. The choices made by device engineers are often guided by algorithmic needs and may at times be

somewhat arbitrary. The numerous stages involved in each device's processing pipeline also make it highly unlikely that different manufacturers would adopt equivalent methods that would yield identical features even with the same input data. In Chapter 4, I examine whether data from various devices yield comparable conclusions about the correlation between these features and key demographic variables.

## 1.2 Acute physiological changes

Human physiology exhibits an impressive ability to maintain balance across a wide range of conditions over time. However, external stressors can lead to disruptions in an individual's underlying physiology that manifest in measurable ways. For example, several acute stressors lead to a transient increase in wearable measured resting heart rate (Alavi et al., 2022). During exercise, heart rate increases, thereby increasing blood flow throughout the body; increased blood flow elevates the rate of delivery of blood-born metabolic agents which helps to compensate for increased metabolic demands (Vatner and Pagani, 1976). Metabolic by-products are also removed at a faster rate. While transient cardiorespiratory changes during exercise are primarily mediated by an increased metabolic demand, several other mechanisms drive such changes in response to an infection. Rather, infection-driven cardiorespiratory changes are regulated by the body's autonomic nervous system, the system by which the body senses and responds to threats (Gordan et al., 2015). The autonomic nervous system senses changes in inflammatory state which manifest as 1) an increase in exogenous inflammatory elements (i.e., antigens), or those stemming from pathogens and their metabolic byproducts during infection, and 2) an increase in the prevalence of endogenous inflammatory elements, or elements released by injured cells (Hannoodee and Nasuruddin, 2024). Immune cells, including monocytes and macrophages, differentially express receptors for these inflammatory elements and produce a host of cytokines and other inflammatory molecules in response to increased inflammatory elements (Parihar et al., 2010). Afferent vagus nerve fibers are sensitive to these inflammatory molecules and are in

part responsible for the cardiorespiratory changes observed during infection (Pavlov and Tracey, 2012). In particular, HRV is mediated via this vagal pathway, often resulting in lower HRVs during an infection (Cooper et al., 2015). HR may also be directly impacted by changes in vagal nerve activation, however, observed increases in HR might be primarily a response to changes in the thermoregulatory set point that occur as a response to infection (Heal et al., 2022). *Increases* in an individual's thermoregulatory set point, also known as a fever, are thought to be mediated by the interaction between inflammatory molecules and thermo-sensitive neurons in the hypothalamus (El-Radhi, 2019). In general, this response is mediated by interleukin-1 (IL-1), however, other inflammatory elements may also lead to an increased thermoregulatory set point (El-Radhi, 2019). IL-1 is one of the many inflammatory molecules produced by macrophages and monocytes in response to inflammatory elements (Madej et al., 2017). Empirically, increases in RR are also commonly observed with increased HR (Heal et al., 2022).

Measurements from wearables seem to be useful for characterizing and detecting these acute physiological changes (Smarr et al., 2020), however, it can be challenging to distinguish between changes caused by exercise and changes caused by acute inflammation events over short timescales. While both exercise and inflammation affect HR, RR, and HRV similarly, the time scale of change differs substantially; exercise-induced changes often revert to baseline within hours but changes due to inflammation manifest over hours to days. As such, in Chapter 2, I examine day-level changes in wearable-measured physiology. This approach reduces noise from minute-to-minute fluctuations and hour-to-hour changes due to physical activity however other stress events such as alcohol consumption still manifest in ways similar to acute inflammation (Alavi et al., 2022). In Chapters 3 and 4, I describe how changes across physiological measurements can be leveraged to develop models to detect these acute inflammation events at scale.

## 1.3 Machine learning in biology

Briefly, the goal of machine learning (ML) is to learn patterns from data. For the sake of this thesis, $X \in \mathbb{R}^D$ and $x \in X$ where $x$ is a $D$ dimensional vector space. I will often refer to a single dimension in $x$ as a feature, however, note that by definition, $x$ being a vector space implies that any given feature might itself be a vector (as may be the case in multivariate time series).

ML is often grouped into either supervised or unsupervised learning, although other paradigms such as reinforcement learning are common (Sarker, 2021). In supervised learning, the goal is to learn a set of labels ($Y$) from data ($X$) whereas in unsupervised learning such labels might not exist, or at least they might not exist in a way that a model is privy to at training time (Sarker, 2021). The models I focus on in Chapters 3 and 4 of this thesis primarily leverage the supervised learning paradigm to train models for the prediction and detection of acute infectious diseases using wearable device data. However, I also use unsupervised learning in Chapter 4 to characterize data across several longitudinal wearable datasets.

Research applying machine learning to biology, broadly construed, has been progressing rapidly, particularly in light of the increasing digitization of biological data (Ghassemi et al., 2020). For example, advances in data-gathering processes and the dissemination of electronic health records (EHRs) have allowed the development of more sophisticated models leveraging these data for clinical prediction tasks (Ghassemi et al., 2020; Pivovarov et al., 2015). Similarly, technological developments have enabled cost-effective and widely accessible genome sequencing and applications to, for example, personalized (i.e., precision) medicine (Brittain et al., 2017). Wearable device data, as studied in this thesis, have enabled large-scale physiological monitoring. While these technologies have increased the digitization of biological data, it is important to consider their use in the context of that data's intended purpose. Historically, models are often trained on data that weren't collected to support model development. Clinical prediction models repurpose EHR data to make personalized predictions, however, EHR systems were initially designed to support billing and later improved care rather than facilitate subsequent analyses

(Ghassemi et al., 2020). Repurposing data for training ML models might bias such models in unintended ways, like driving them to make predictions that human practitioners make rather than some ideal (Gianfrancesco et al., 2018). The wearable data I employed in this thesis were first and foremost intended to inform individual consumers and were not necessarily designed for large cross-sectional studies as I do in this thesis. However, efforts using wearable data suggest they can provide novel insights into human physiology. In Chapter 4, I examine whether simple assumptions about wearable device data lead to consistent conclusions across large-scale wearable device datasets.

Implicitly, supervised learning models leverage data to make predictions. Models that predict continuous scalar values $Y$ from $X$ are often called regression models and models that predict one or more of $C$ discrete, so-called, classes are often called classifiers (Sarker, 2021). The models I develop throughout this thesis are primarily of the classification type and aim to predict the presence or absence of a particular acute physiological change (i.e., fever symptoms, flu symptoms, and viral infection). As ML models learn patterns from data, their performance is largely determined by that data (Lei et al., 2023). Choices made by practitioners in how they *featurize* untransformed data can greatly impact model performance. For example, training a model on untransformed feature vectors $x$ as they are computed by each wearable device (i.e., heart rate) would lead to a model that learns an optimal decision boundary (minimizes some loss function $L$) against some population average for that feature. Inclusion of demographic information (i.e., age, biological sex, ethnicity) as features might allow a model to condition its decision boundary (*personalize* the model) over these demographic features and leverage, for example, the fact that African American participants tend to exhibit higher HRs than Caucasian participants to adjust its decision boundary accordingly (Liu et al., 1989). Practically speaking, there are reasons to be wary of models that require such demographic information; ethnicity, at least as it is commonly categorized, might not neatly compartmentalize humans in a biologically useful way (Jackson, 1992), and personalized models often perform worse on certain subgroups (Suriyakumar et al., 2023). More important in the case of wearable

data, treating such untransformed features as unrelated examples violates the assumption of independent and identically distributed samples (Darrell et al., 2015). Untransformed features also lose the context of much of the important time-varying information that characterizes human physiology. In this work, I leverage the ability to compare an individual to themselves over time to improve the featurization of wearable device data. As I outline further in Chapters 3 and 4, featurization optimization substantially improves model performance.

### 1.3.1 Noisy labels

The labels (the $Y$ in supervised learning) used to train models based on human biological data might exhibit biases (Karimi et al., 2020). In cases where humans are the primary source of a model's ground truth, collected for instance via an optional survey, several known issues permeate from labels to that model's predictions as humans are notorious for bias in several domains. The problem of biased labeling is commonly referred to as *label noise* (Nagaraj et al., 2024). While the name label "noise" implies an element of randomness, it is not uncommon that differences between provided labels and the ideal labels are systematic; label noise arises when annotators lack the requisite expertise to effectively label a set of data (Nagaraj et al., 2024). In EHR data, a clinician's ability to accurately diagnose patients might degrade when persistently interrupted (Westbrook et al., 2010). In Chapters 2, 3, and 4, I rely on labels provided by participants in each study. In particular, I rely on participants' subjective experience of certain symptoms. Certain biases are known to impact such subjective labels; demographic factors are significant predictors of the likelihood that an individual reports certain symptoms (Kroenke and Spitzer, 1998) and individuals often exhibit recall bias (Althubaiti, 2016). While label noise can present challenges in other domains like image classification, these data can be re-labeled after the fact and even leverage consensus labels (i.e., majority voting) across ad hoc labelers (Ipeirotis et al., 2010). In contrast, subjective labeling of wearable time series is a unique challenge, where only the individuals themselves can provide ground truth labels (Yamagata et al., 2023). Secondary annotators might not be capable of reliably labeling these data.

While the validity of any human's label of a particular instance of biological data (the $x$ in supervised learning) can impact a model's performance, how researchers choose to create an ML task based on these data also drives both how well any particular model appears to perform on a particular test set, as in its accuracy or area under the receiver operator curve (AUROC), as well as its performance in real-world deployment scenarios. In the case of clinical models trained on EHR data, the choice of ground truth can be more important than the specific model used (Cohen et al., 2024). Choices of task definitions should ideally optimize an objective directly relevant to how a model would prove useful in deployment. As I outline in Chapter 4, there is substantial variability in how the community of researchers developing models to detect the onset of acute illnesses using wearable device data define an important concept: the onset of acute illness. Substantial differences in the definition of these concepts can make it challenging to compare the performance of different models across implementations and datasets. As I demonstrate in Chapter 3, for a set of individuals, it might be possible to leverage an individual's wearable data to predict whether that individual will present fever symptomology prior to the onset of their subjective experience of that symptom. An open issue within this field is how a practitioner might systematize such predictions in a useful manner. I argue in Chapter 4 that specifically within the field of acute illness detection using wearables, some unified standard might be agreed upon and I provide a first look at systematizing such processes.

Human data are often subject to random and structured variance. First, note that data-generating processes themselves might bias information structures. In the case of EHR data, certain clinical tests might only be ordered when it is likely that a patient has a particular disease (Realdi et al., 2008; Ghassemi et al., 2020). A feature based on such a clinical test might only be informative in whether it exists in the EHR or not (i.e., is missing). A feature where its presence or absence is totally random is commonly referred to as missing completely at random (MCAR) (Wells et al., 2013; Ghassemi et al., 2020). The missingness of MCAR features is non-informative (Wells et al., 2013; Ghassemi et al., 2020). In the case of wearable device data, the time when a user takes off their device to charge might be MCAR or it might be missing

not at random (MNAR) (Wells et al., 2013; Ghassemi et al., 2020). Practitioners can leverage knowledge about the missingness structure of certain features to design datasets that can be used to develop models that are more robust to feature noise. For example, simply removing training examples with missing MCAR features is thought to be more justified than removing examples with missing MNAR features. Instead, MNAR features might need to be imputed in some way. Throughout this thesis, I leverage assumptions about the missingness structure of wearable data to construct training datasets.

The remainder of this thesis is comprised of four additional chapters. In Chapter 2, I use wearable device data to characterize acute physiological changes around the administration of the COVID-19 vaccine and self-reported fever onset. In Chapter 3, I develop a model for detecting the physiological manifestations of self-reported fevers and describe its performance in a large population. In Chapter 4, I compare several large-scale, longitudinal wearable device datasets and examine the generalizability of acute illness onset detection algorithms across these datasets. Finally, in Chapter 5 I provide a discussion of future directions for this research.

# Chapter 2

# Characterizing acute physiological changes using wearable device data

Wearable device data exhibit potential in both specific and general health surveillance contexts. First, I demonstrate, along with my co-authors, that physiological changes recorded by the Oura Ring post-COVID-19 vaccination, such as increases in dermal temperature and heart rate, correlate significantly with higher productions of SARS-CoV-2 neutralizing antibodies. Notably, greater temperature deviations were the strongest predictors of enhanced antibody responses. This study involved 1,179 participants who wore the device and had their antibody levels tested during the U.S. vaccination rollout. Conversely, the second study explored the broader application of wearable technology in detecting fever episodes as part of syndromic surveillance, involving a larger sample of 63,153 participants. Using the same data from the Oura Ring, including dermal temperature and heart rate from days surrounding reported fevers, I developed a highly performant tree-based classifier. This indicates a promising avenue for real-time disease surveillance and prevalence monitoring during infectious disease outbreaks, albeit with challenges in false positive rates and sensitivity. Together, these studies underscore the versatility and utility of wearable devices in both targeted and broad public health applications.

## 2.1 Physiological changes in response to vaccination

COVID-19 vaccines have been highly effective in preventing severe disease, demonstrating reductions in risk by about 90% (Heath Paul T. et al., 2021; Baden Lindsey R. et al., 2021), though they do not completely eliminate the risk of severe disease and concerns exist about waning protection over time (Pouwels et al., 2021; Feng et al., 2021). The effectiveness of these vaccines correlates strongly with levels of antibodies against the SARS-CoV-2 spike protein which is responsible for viral neutralization (Feng et al., 2021). Variants like Omicron, which show reduced sensitivity to these antibodies, highlight the importance of high antibody levels for optimal protection (Wilhelm et al., 2022). Prior research efforts explored whether modifiable factors like sleep duration (Spiegel et al., 2002; Lange et al., 3 10) and physiological responses to vaccination, such as fever and fatigue, influence antibody levels (Debes et al., 2021). This study sought to use data from the Oura Ring to monitor metrics like dermal temperature and heart rate variability to identify predictors of antibody responses by measuring antibodies to the SARS-CoV-2 receptor binding domain (RBD), a key marker of immune protection. This work helps determine whether physiological responses and lifestyle factors impact vaccine-induced immunity and contributes to evidence from existing studies which suggest a correlation between vaccine side effects and antibody production.

### 2.1.1 Materials and Methods

We initiated the second TemPredict study in December of 2020 to assess whether an algorithm derived from physiological metrics collected by an off-the-shelf wearable device (Oura Ring) could be used to detect COVID-19 infection in real-time. An additional aim of this study was to assess whether data from this device could predict antibody response quantified as antibodies to SARS-CoV-2 spike protein receptor binding domain (RBD), which is the focus of the current report.

## Study Participants

*Recruitment.* We recruited participants residing in the United States who already possessed Oura Rings by sending them email invitations. These email invitations included a link to an online consent survey. We also recruited participants who worked at participating sites (e.g., teachers, firefighters, and other first responders) by enlisting leadership at these sites to assist in recruitment. We mailed these sites recruitment materials, including study flyers and Oura Ring sizing kits, which contained plastic rings for prospective participants to try on to determine their size. We provided Oura Rings to interested individuals at these sites after they provided their size information to study coordinators.

*Eligibility and Consent.* Eligible participants were at least 18 years of age, possessed a smartphone that could pair with their Oura Ring, resided in the United States, did not previously have COVID-19 infection (verified through laboratory testing during enrollment), and could communicate in English. For this analysis, of the 2055 participants who completed the overall study, we first excluded participants who had a positive SARS-CoV-2 nucleocapsid antibody test at the end of the study, indicating COVID-19 infection during the study period (n = 56). We then excluded participants who were not fully vaccinated at least 7 nights prior to their final blood draw (n = 715). We then excluded participants who did not have at least 7 nights of physiological data within the timeframe used to develop the pre-vaccination baseline period (night -14 to night -4 prior to first vaccination or who lacked data for at least one night adjacent to vaccination; n = 105).

The University of California San Francisco (UCSF) Institutional Review Board (IRB, IRB# 20-30408) and the U.S. Department of Defense (DOD) Human Research Protections Office (HRPO, HRPO# E01877.1a) approved of all study activities, and all research was performed in accordance with relevant guidelines and regulations. All participants provided electronic (written) informed consent, and this research was conducted according to the principles expressed in the Declaration of Helsinki. Participants to whom we provided Oura Rings kept the devices

following their participation; we did not otherwise compensate participants for participation.

## Measures

*Questionnaires:* Beginning in December 2020, participants completed several online surveys. They first completed a baseline survey that collected demographic and health information. They also completed daily and monthly surveys on which they reported COVID-19 symptoms, COVID-19 diagnosis, and COVID-19 exposures. Within these surveys, participants also reported whether they had been vaccinated against COVID-19, and if so, which vaccine they received (Pfizer-BioNTech, Moderna-NIAID, or Johnson & Johnson-Janssen) as well as their injection dates. After participants reported on vaccine type and dates of vaccine injections, their surveys were customized such that they were not asked these questions in duplicate on future dates.

*Antibody testing.* We tested participants for antibodies to the SARS-CoV-2 nucleocapsid protein (Test# 164068, LabCorp, Inc.) during enrollment (December 2020 through early April 2021) and at the end of their participation (April and May 2021). The SARS-CoV-2 nucleocapsid antibody test becomes positive following COVID-19 infection, but vaccination does not cause individuals to generate antibodies to this part of the virus. Participants were required to have a negative nucleocapsid antibody test at enrollment as we excluded participants with evidence of prior COVID-19 infection. At the end of the study period (late April and May 2021), we also tested participants for antibodies to the SARS-CoV-2 RBD with the LabCorp Semi-Quantitative Total Antibody, Spike assay (Test# 164090, LabCorp, Inc.), which used the Roche Elecsys Anti-SARS-CoV-2 S assay performed on a COBAS e602 module (Roche Diagnostics, 2022). The specificity and sensitivity ($\geq$14 days post-PCR diagnosis of COVID-19 infection) for the Elecsys Anti-SARS-CoV-2 S immunoassay is reported to be 99.95% (95% CI: 99.87–99.99) and 97.92% (95% CI: 95.21–99.32), respectively (Riester et al., 2021). The dynamic range reported for this assay during the time when most of the study assays were performed was from 0.4 IU/mL to 2500 IU/mL, with a clinical cut-off value for positive results of 0.8 U/mL. Prior to 3 May 2021, LabCorp reported results using an upper detection limit of 250 IU/mL, after which

LabCorp changed assay procedures to quantitate antibody levels up to 2500 IU/mL. Fifty-seven participants completed the RBD antibody test before 3 May 2021 and had a result of ">250 IU/mL".

*Wearable device data (device-generated metrics).* Participants wore the Oura Ring (Generation 2), a commercially available wearable sensor device (Oura Health, Oulu, Finland), on a finger of their choosing. The Oura Ring connects to the Oura App (available from the Google Play Store and the Apple App Store) via Bluetooth. Users can wear the ring continuously in both wet and dry environments.

The Oura Ring generates physiological metrics by aggregating data gathered from on-device sensors. These high-resolution metrics are transformed into summary metrics before transmission to a smartphone app. These device-generated metrics include nightly summary variables of dermal temperature deviations, resting heart rate (HR), resting heart rate variability (HRV), and respiratory rate (RR). The Oura Ring Gen2 assesses HR, HRV, and RR from a photoplethysmogram (PPG) signal generated at 250 Hz. The Oura ring calculates HR, HRV, and RR from inter-beat intervals (IBI), which the Oura Ring only generates during periods of sleep. The Oura Ring calculates HRV in the form of the root mean square of the successive differences (RMSSD). Tri-axial accelerometers estimate activity metrics as metabolic equivalents (MET) reported at 10–60 Hz during both sleep and wake periods, and sleep stages at 5 min resolution. The Oura Ring assesses temperature by using a negative temperature coefficient (NTC) thermistor (resolution of 0.07 °C) on the internal surface of the ring. The sensor registers dermal temperature readings from the palm side of the finger base every 60 s. The temperature deviation metric is computed as the difference between a user's average overnight temperature and their longer-term baseline, calculated using a rolling window roughly equal to the prior two months. The Oura Ring also outputs sleep metrics that include minutes of light sleep (non-rapid-eye movement [NREM] stages 1 and 2), deep sleep (NREM stages 3 and 4), rapid eye movement (REM) sleep, and total sleep time. We examined these metrics (temperature deviation, HR, HRV, RR, REM sleep duration, deep sleep duration, and total sleep duration) in the present

analyses.

*Vaccination.* Participants reported the dates on which they received injections of one of the three vaccines available in the United States (Pfizer-BioNTech, Moderna-NIAID, or Johnson & Johnson-Janssen) between December 2020 and May 2021.

**Analytic Plan**

*Outcome.* For correlation analyses, we treated values of ">2500 IU/mL" as 2500 IU/mL (n = 474). We omitted participants who received a value of ">250 IU/mL;" (n = 51) from primary analyses. However, we reported analyses including these values as 250 in supplementary results.

*Predictors.* We used device-generated values of each metric two nights before and three nights after each injection (nights -2, -1, 0, 1, 2, where 0 represents the night of the day when vaccination occurred). We established a pre-vaccination baseline period for each participant from 14 nights to 4 nights prior to the first vaccine injection (night -14 to night -4). We calculated values of each device-generated metric adjusting for this pre-vaccination baseline period by converting each physiological metric to a z-score using participants' respective individual means and standard deviations from the pre-vaccination baseline period (this transformation allows analysis of relative change, without the need for individual device calibration). We examined device-generated metrics from all nights surrounding each injection (-2, -1, 0, 1, 2) and device-generated metrics adjusted for the pre-vaccination baseline period as predictors of RBD antibody responses. Notably, all device-generated metrics reflect values solely from the prior night, except the temperature deviation metric (computed as the difference between the prior night's value as a deviation from an average derived of the prior two months). The temperature deviation metric adjusted for the pre-vaccination baseline period, therefore, reflects a difference in two deviation metrics: the difference between a participant's (1) deviation on a particular night surrounding injection and (2) average deviation during the pre-vaccination baseline period.

If a participant was missing device-generated metrics from a particular night, we did not

include them in analyses for that night. We excluded participants who had a positive SARS-CoV-2 nucleocapsid antibody test at the end of the study, who did not receive their second injection (Moderna-NIAID and Pfizer-BioNTech) at least 7 nights prior to their final blood draw, who did not have at least 7 nights of physiological data within pre-vaccination baseline period (night -14 to night -4), and who had a threshold value of "$>$250 IU/mL" on the outcome.

*Statistical analyses.* We conducted analyses separately for each type of vaccine (Johnson & Johnson-Janssen, Moderna-NIAID, Pfizer-BioNTech). We also analyzed the data from both mRNA vaccines combined (Moderna-NIAID and Pfizer-BioNTech). First, we conducted Spearman rank-order correlations between RBD antibody responses and device-generated metrics, before and after adjusting for the pre-vaccination baseline period. We also examined correlations between metrics assessed during the pre-vaccination baseline period and RBD antibody responses.

Second, we used the results of the bivariate correlational analyses to inform variable selection for multivariate regression models that assessed which device-generated metrics independently predicted RBD antibody responses. Based on results from the Spearman correlations, we combined data from the mRNA vaccine recipients (Pfizer-BioNTech and Moderna-NIAID) from night 0 after the second injection to predict RBD antibody responses from device-generated metrics before and after adjusting for the pre-vaccination baseline period. We included device-generated metrics that had associations in Spearman correlation analyses (in the combined mRNA sample) before and/or after adjusting for the pre-vaccination baseline period with the RBD antibody responses with p-values $< 0.1$. Due to the proportion of observations with right-censored values of the outcome variable, we adopted a semi-parametric approach using Cox regression models (Vittinghoff and McCulloch, 2007), using the RBD antibody result as the dependent variable. Because these analyses used Cox regression to assess differences in antibody levels (rather than time to event typically used in Cox regression), we reported coefficients rather than hazard ratios usually reported with Cox analyses. Coefficients offer insight into the direction and magnitude of associations between the RBD antibody responses and device-generated metrics

(Therneau and Grambsch, 2000). We found neither strong nor linear effects of time on antibody titer during the study period. As a result, we did not include temporal parameters in the Cox regression models.

### 2.1.2 Results

We enrolled 2392 participants in the second TemPredict study (Figure 2.1 and Table 2.1). After excluding participants who did not receive their second injection (Moderna-NIAID and Pfizer-BioNTech) at least 7 nights prior to their final blood draw, who did not have at least 7 nights of physiological data within the pre-vaccination baseline period (night -14 to night -4), who had a value of "$>$250 IU/mL" on the RBD antibody assay, or who had a positive value on their second nucleocapsid antibody test, there were 1179 participants eligible for this analysis. Of these participants, 107 received Johnson & Johnson-Janssen COVID-19 vaccine, 366 received the Moderna-NIAID vaccine, and 706 received the Pfizer-BioNTech vaccine (Figure 2.1). Of the 1179 participants included in this analysis, 474 had a result of "$>$2500 IU/mL" on the RBD antibody test, indicating substantial right censoring (40.2%) of the data. Four participants in the analysis dataset had a left-censored RBD value ("$<$0.4 IU/mL"). Participants in the analytic sample obtained the RBD test an average of 38 days (SD: 30 days) after the final vaccine injection. By vaccine types, the mean (SD) days from final vaccination to RBD testing were: Moderna-NIAID 35 (24), Pfizer-BioNTech 40 (34), and Johnson & Johnson-Janssen 39 (14).

#### Spearman Correlations

Changes in device-generated metrics during night 0 (the night immediately following the second injection) tended to show a stronger pattern of associations with RBD antibody responses than these metrics the night immediately following the first injection (Table 2.2 and Figure 2.2). In some cases, these associations were also evident the following night (night 1) after the second injection.

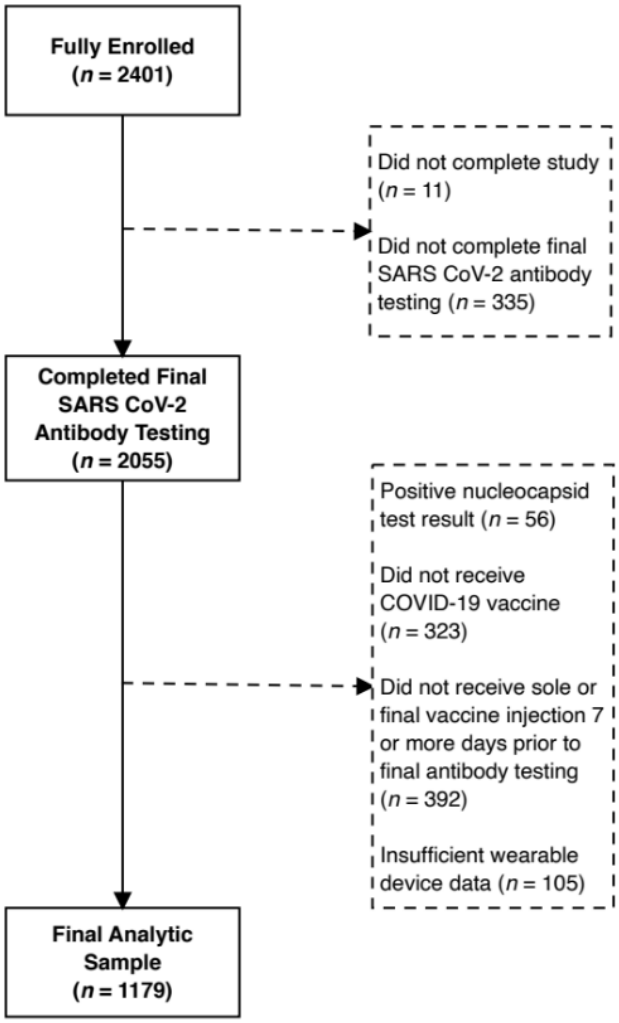


**Figure 2.1.** Participant flow through the study.

**Table 2.1.** Participant characteristics.

| | **Pfizer-BioNTech** | **Moderna-NIAID** | **Johnson & Johnson-Janssen** | **Overall** |
|---|---|---|---|---|
| N | 706 | 366 | 107 | 1179 |
| Age (M, SD) | 50.4 (11.4) | 52.8 (12.0) | 49.5 (9.9) | 51.0 (11.5) |
| Biological sex (N, %) | | | | |
| Male | 334 (47.3) | 166 (45.4) | 53 (49.5) | 553 (46.9) |
| Female | 371 (52.5) | 200 (54.6) | 54 (50.5) | 625 (53.0) |
| Intersex | 1 (0.1) | 0 (0) | 0 (0) | 1 (0.1) |
| Race (N, %) | | | | |
| African | 1 (0.1) | 1 (0.3) | 0 (0) | 2 (0.2) |
| African American | 9 (1.3) | 6 (1.6) | 2 (1.9) | 17 (1.4) |
| Caribbean | 1 (0.1) | 1 (0.3) | 0 (0) | 2 (0.2) |
| Caucasian/White | 603 (85.4) | 325 (88.8) | 95 (88.8) | 1023 (86.8) |
| East Asian | 37 (5.2) | 9 (2.5) | 4 (3.7) | 50 (4.2) |
| Middle Eastern | 5 (0.7) | 3 (0.8) | 1 (0.9) | 9 (0.8) |
| Native American/ Native Alaskan | 3 (0.4) | 0 (0) | 0 (0) | 3 (0.3) |
| Native Hawaiian / Other Pacific Islander | 0 (0) | 0 (0) | 1 (0.9) | 1 (0.1) |
| South Asian | 16 (2.3) | 3 (0.8) | 1 (0.9) | 20 (1.7) |
| More than 1 race | 24 (3.4) | 15 (4.1) | 2 (1.9) | 41 (3.5) |
| Prefer not to answer / Unknown | 7 (1.0) | 3 (0.8) | 1 (0.9) | 11 (0.9) |
| Hispanic/Latinx (N, %) | | | | |
| Yes | 30 (4.2) | 17 (4.6) | 7 (6.5) | 54 (4.6) |
| No | 671 (95.0) | 347 (94.8) | 100 (93.5) | 1118 (94.8) |
| Don't Know / Not Sure | 4 (0.6) | 1 (0.3) | 0 (0) | 5 (0.4) |
| Prefer not to answer | 1 (0.1) | 1 (0.3) | 0 (0) | 2 (0.2) |
| Education (N, %) | | | | |
| Less than a high school diploma | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| High school diploma or GED | 5 (0.7) | 3 (0.8) | 3 (2.8) | 11 (0.9) |
| Some college | 40 (5.7) | 27 (7.4) | 10 (9.3) | 77 (6.5) |
| Associate Degree (e.g., AA, AS) | 27 (3.8) | 13 (3.6) | 6 (5.6) | 46 (3.9) |
| Bachelor's Degree (e.g., BA, BS) | 270 (38.2) | 131 (35.8) | 40 (37.4) | 441 (37.4) |
| Master's Degree (e.g., MA, MS) | 223 (31.6) | 119 (32.5) | 38 (35.5) | 380 (32.2) |
| Advanced Degree (e.g., PhD, EdD, MD, JD) | 141 (20.0) | 73 (19.9) | 10 (9.3) | 224 (19.0) |
| RBD Value (N, %) | | | | |
| Value of 0 to 2499 IU/ml | 458 (64.9) | 89 (24.3) | 107 (100.0) | 654 (55.5) |
| Value of ">250 IU/ml" | 33 (4.7) | 18 (4.9) | 0 (0) | 51 (4.3) |
| Value of ">2500 IU/ml" | 215 (30.5) | 259 (70.8) | 0 (0) | 474 (40.2) |
| RBD Value (Median, IQR) | | | | |
| Value of 0 to 2499 IU/ml | 1613.0 [778.4, 2500.0] | 2500.0 [2477.0, 2500.0] | 19.1 [6.1,50.1] | 1956.5 [753.3, 2500.0] |

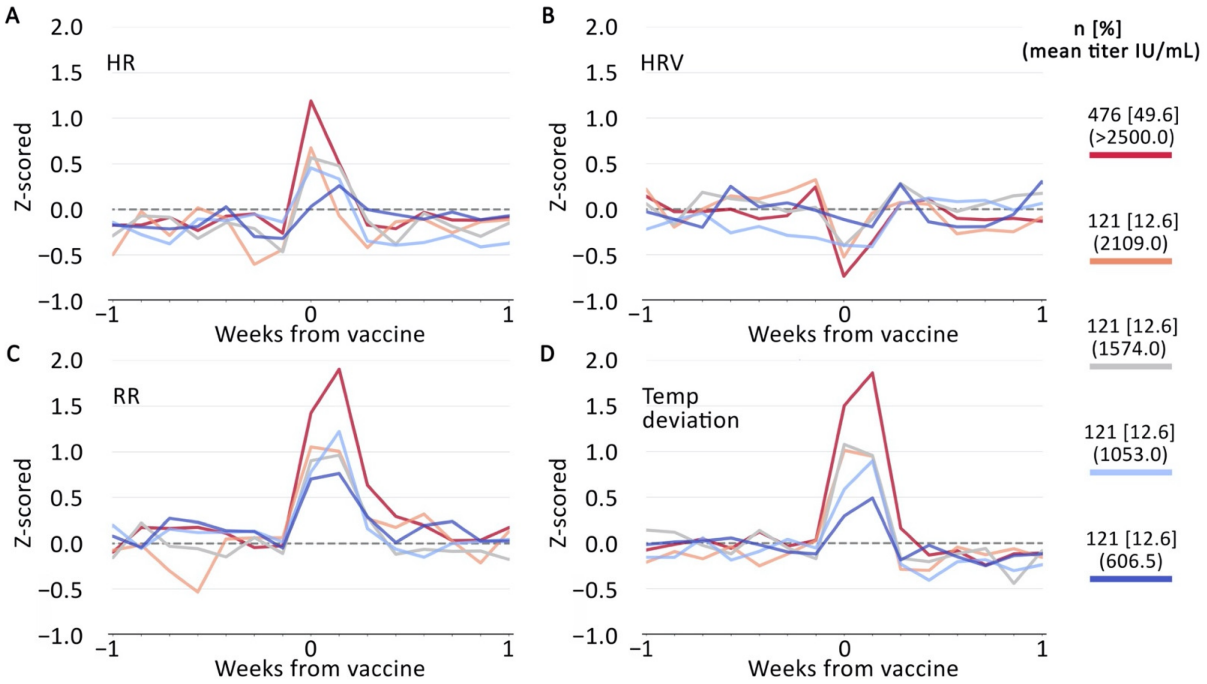


**Figure 2.2.** Plots depicting (A) changes in heart rate (HR); (B) heart rate variability (HRV); (C) respiratory rate (RR); and (D) temperature deviation the nights surrounding the second injection for Pfizer-BioNTech and Moderna-NAIAD vaccine recipients, combined. Values are z-scored for participants' pre-vaccination baseline period (see Materials and Methods).

**Table 2.2.** Spearman rank-order correlations between RBD antibody responses and device-generated metrics on nights before and after each vaccine injection.

| Night Relative to Injection | Metric | Injection 1 J&J | | Injection 1 Moderna | | Injection 1 Pfizer | | Injection 2 Moderna | | Injection 2 Pfizer | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **rho** | *P* | **rho** | *P* | **rho** | *P* | **rho** | *P* | **rho** | *P* |
| -2 | Sleep Duration | -0.127 | 0.207 | -0.031 | 0.571 | 0.018 | 0.649 | 0.069 | 0.218 | -0.023 | 0.564 |
| | REM Sleep | 0.026 | 0.799 | -0.036 | 0.518 | **0.104** | **0.009** | 0.008 | 0.886 | 0.061 | 0.126 |
| | Deep Sleep | -0.139 | 0.166 | -0.024 | 0.657 | 0.01 | 0.808 | 0.07 | 0.214 | 0.049 | 0.219 |
| | HRV (RMSSD) | -0.119 | 0.236 | 0.013 | 0.812 | -0.046 | 0.251 | 0.007 | 0.908 | 0.025 | 0.536 |
| | HR | -0.038 | 0.706 | 0.033 | 0.547 | **0.138** | **0** | 0.031 | 0.583 | 0.063 | 0.117 |
| | RR | 0.072 | 0.477 | -0.048 | 0.381 | -0.067 | 0.092 | **-0.13** | **0.02** | -0.045 | 0.26 |
| | Temp Deviation | -0.028 | 0.784 | -0.012 | 0.83 | 0.052 | 0.193 | 0.043 | 0.439 | -0.005 | 0.891 |
| -1 | Sleep Duration | 0.036 | 0.723 | -0.022 | 0.683 | -0.044 | 0.27 | -0.059 | 0.289 | -0.068 | 0.088 |
| | REM Sleep | 0.018 | 0.856 | -0.062 | 0.255 | 0.033 | 0.403 | -0.086 | 0.126 | 0.001 | 0.976 |
| | Deep Sleep | 0.012 | 0.908 | -0.048 | 0.375 | 0.047 | 0.236 | 0.005 | 0.933 | 0.048 | 0.235 |
| | HRV (RMSSD) | -0.081 | 0.421 | 0.049 | 0.366 | 0.001 | 0.974 | 0.028 | 0.614 | 0.044 | 0.275 |
| | HR | 0.056 | 0.574 | 0.021 | 0.704 | **0.103** | **0.009** | 0.02 | 0.726 | **0.093** | **0.02** |
| | RR | 0.122 | 0.223 | -0.047 | 0.387 | **-0.094** | **0.018** | -0.048 | 0.391 | **-0.079** | **0.048** |
| | Temp Deviation | -0.047 | 0.64 | -0.023 | 0.67 | 0.002 | 0.951 | 0.053 | 0.342 | -0.023 | 0.568 |
| 0 | Sleep Duration | 0.125 | 0.211 | 0.066 | 0.235 | **-0.08** | **0.044** | -0.003 | 0.96 | -0.033 | 0.408 |
| | REM Sleep | 0.054 | 0.593 | 0.045 | 0.413 | -0.002 | 0.968 | 0.024 | 0.659 | 0.033 | 0.407 |
| | Deep Sleep | -0.162 | 0.106 | -0.01 | 0.86 | 0.005 | 0.899 | -0.025 | 0.654 | -0.058 | 0.148 |
| | HRV (RMSSD) | -0.047 | 0.638 | 0.011 | 0.838 | 0.006 | 0.873 | -0.056 | 0.312 | **-0.092** | **0.022** |
| | HR | 0.125 | 0.213 | 0.008 | 0.885 | **0.101** | **0.01** | **0.138** | **0.012** | **0.176** | **0** |
| | RR | 0.123 | 0.222 | -0.053 | 0.34 | -0.058 | 0.141 | -0.021 | 0.711 | -0.004 | 0.925 |
| | Temp Deviation | 0.155 | 0.122 | -0.009 | 0.87 | -0.003 | 0.935 | **0.123** | **0.026** | **0.152** | **0** |
| 1 | Sleep Duration | **-0.202** | **0.044** | -0.053 | 0.334 | -0.002 | 0.957 | 0.069 | 0.221 | -0.009 | 0.823 |
| | REM Sleep | -0.056 | 0.58 | 0.023 | 0.674 | 0.045 | 0.258 | 0.068 | 0.231 | **0.082** | **0.041** |
| | Deep Sleep | -0.149 | 0.138 | -0.006 | 0.912 | 0.042 | 0.29 | -0.09 | 0.11 | 0.026 | 0.522 |
| | HRV (RMSSD) | -0.073 | 0.471 | 0.003 | 0.962 | -0.012 | 0.765 | 0.011 | 0.842 | -0.016 | 0.696 |
| | HR | 0.049 | 0.631 | 0.002 | 0.968 | **0.087** | **0.028** | 0.053 | 0.352 | **0.116** | **0.004** |
| | RR | 0.157 | 0.119 | -0.03 | 0.585 | -0.066 | 0.098 | -0.007 | 0.899 | 0.049 | 0.229 |
| | Temp Deviation | 0.093 | 0.356 | 0.055 | 0.316 | -0.03 | 0.451 | **0.136** | **0.016** | **0.186** | **0** |
| 2 | Sleep Duration | 0.046 | 0.646 | -0.019 | 0.735 | -0.016 | 0.684 | -0.061 | 0.282 | **-0.106** | **0.008** |
| | REM Sleep | 0.019 | 0.853 | -0.026 | 0.641 | 0.032 | 0.425 | -0.038 | 0.497 | -0.001 | 0.974 |
| | Deep Sleep | 0.026 | 0.795 | -0.039 | 0.478 | -0.006 | 0.87 | 0.012 | 0.832 | -0.013 | 0.739 |
| | HRV (RMSSD) | -0.074 | 0.465 | 0.038 | 0.491 | 0.005 | 0.891 | 0.004 | 0.942 | -0.041 | 0.312 |
| | HR | 0.108 | 0.282 | 0.033 | 0.552 | 0.089 | **0.025** | 0.041 | 0.467 | **0.1** | **0.013** |
| | RR | 0.144 | 0.152 | 0 | 0.997 | -0.09 | **0.023** | -0.048 | 0.394 | -0.027 | 0.496 |
| | Temp Deviation | 0.187 | 0.062 | 0.001 | 0.987 | -0.019 | 0.641 | 0.081 | 0.153 | 0.021 | 0.594 |

**Device-Generated Metrics**

Greater HR and temp deviation on night 0 after the second injection for both the Moderna-NIAID (HR: rho = 0.138, p = 0.012; temp deviation: rho = 0.123, p = 0.026) and Pfizer-BioNTech (HR: rho = 0.176, $p < 0.001$; temp deviation: rho = 0.152, $p < 0.001$) were associated with greater RBD antibody responses (Table 2.2). Additionally, on night 0 after the second injection, lower HRV values were associated with greater RBD antibody responses for Pfizer-BioNTech (HRV: rho = -0.092, p = 0.022), and these associations were in the same direction, but not statistically significant, for Moderna-NIAID (HRV: rho = -0.056, p = 0.312). The associations between RBD antibody responses and HR (Pfizer-BioNTech only) and temp deviation (Pfizer-BioNTech and Moderna-NAIAD) were also evident the following night (night 1) after the second injection. Correlations between device-generated metrics and antibody responses were in the same direction and often had similar rho values for participants who received the Johnson & Johnson-Janssen vaccine, but none of these correlations were statistically significant in the smaller group that received this vaccine.

When analyzing the two mRNA vaccines in combination (Moderna-NIAID and Pfizer-BioNTech) yielded similar associations between device-generated metrics and RBD antibody responses (Table 2.3). HR, temp deviation, and HRV from night 0 after the second injection were significantly associated with RBD antibody responses (HR: rho = 0.197, $p < 0.001$; temp deviation: rho = 0.238, $p < 0.001$; HRV: rho = -0.118, $p < 0.001$). In analyses combining participants who received either of the two mRNA vaccines, there was a statistically significant inverse correlation between deep sleep on night 0 after the second injection and RBD antibody responses (Deep: rho = -0.079, p = 0.014). The associations between RBD antibody responses and HR, RR, and temp deviation were also evident the following night (night 1) after the second injection.

**Table 2.3.** Spearman rank order correlations between RBD antibody responses and device-generated metrics before and after adjusting for the pre-vaccination baseline period on nights before and after injections for Moderna-NIAID and Pfizer-BioNTech vaccine recipients, combined.

| | | | **Device-generated Metric** | | | | **Adjusted by Baseline period** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Metric** | **Injection 1** | | **Injection 2** | | **Injection 1** | | **Injection 2** | |
| | | | **rho** | *P* | **rho** | *P* | **rho** | *P* | **rho** | *P* |
| **Night Relative to Injection** | -2 | Sleep Duration | -0.001 | 0.982 | 0.021 | 0.519 | **0.001** | **0.988** | 0.062 | 0.057 |
| | | REM Sleep | 0.056 | 0.079 | **0.059** | **0.071** | 0.028 | 0.38 | 0.038 | 0.245 |
| | | Deep Sleep | -0.008 | 0.808 | 0.042 | 0.194 | -0.004 | 0.901 | 0.04 | 0.221 |
| | | HRV (RMSSD) | -0.031 | 0.338 | **-0.007** | **0.819** | **-0.065** | **0.043** | 0.004 | 0.904 |
| | | HR | **0.1** | **0.002** | **0.051** | **0.12** | **0.081** | **0.012** | -0.022 | 0.496 |
| | | RR | -0.041 | 0.197 | **-0.071** | **0.03** | **0.058** | **0.071** | -0.011 | 0.735 |
| | | Temp Deviation | 0.017 | 0.603 | 0.001 | 0.967 | 0.044 | 0.167 | 0.023 | 0.488 |
| | -1 | Sleep Duration | -0.011 | 0.729 | -0.042 | 0.199 | -0.01 | 0.764 | -0.023 | 0.488 |
| | | REM Sleep | 0.036 | 0.264 | -0.015 | 0.648 | 0.008 | 0.798 | -0.05 | 0.123 |
| | | Deep Sleep | 0.017 | 0.594 | 0.031 | 0.34 | 0.056 | 0.079 | 0.024 | 0.463 |
| | | HRV (RMSSD) | -0.001 | 0.981 | 0.043 | 0.186 | 0.011 | 0.741 | **0.074** | **0.024** |
| | | HR | **0.07** | **0.029** | **0.052** | **0.114** | -0.009 | 0.773 | **-0.011** | **0.746** |
| | | RR | -0.056 | 0.08 | **-0.064** | **0.048** | 0.009 | 0.77 | **-0.026** | **0.43** |
| | | Temp Deviation | 0.008 | 0.813 | -0.015 | 0.644 | 0.036 | 0.264 | 0.018 | 0.591 |
| | 0 | Sleep Duration | -0.023 | 0.476 | **-0.012** | **0.704** | -0.018 | 0.565 | 0.024 | 0.462 |
| | | REM Sleep | 0.01 | 0.755 | 0.031 | 0.343 | -0.024 | 0.46 | 0.022 | 0.497 |
| | | Deep Sleep | -0.025 | 0.43 | **-0.079** | **0.014** | -0.014 | 0.667 | **-0.126** | **0** |
| | | HRV (RMSSD) | -0.01 | 0.749 | **-0.118** | **0** | **-0.007** | **0.816** | **-0.199** | **0** |
| | | HR | **0.07** | **0.029** | **0.197** | **0** | **0.013** | **0.695** | **0.208** | **0** |
| | | RR | -0.039 | 0.221 | 0.038 | 0.241 | **0.069** | **0.03** | **0.177** | **0** |
| | | Temp Deviation | 0.006 | 0.85 | **0.238** | **0** | **0.023** | **0.471** | **0.244** | **0** |
| | 1 | Sleep Duration | -0.018 | 0.573 | **0.065** | **0.047** | -0.016 | 0.625 | **0.125** | **0** |
| | | REM Sleep | 0.034 | 0.289 | **0.081** | **0.014** | **0.01** | **0.757** | **0.084** | **0.011** |
| | | Deep Sleep | 0.014 | 0.655 | -0.012 | 0.71 | **0.028** | **0.381** | -0.021 | 0.522 |
| | | HRV (RMSSD) | -0.032 | 0.317 | -0.029 | 0.385 | **-0.08** | **0.013** | **-0.074** | **0.024** |
| | | HR | **0.072** | **0.024** | **0.1** | **0.002** | 0.035 | 0.281 | **0.083** | **0.012** |
| | | RR | -0.015 | 0.643 | **0.067** | **0.042** | **0.097** | **0.002** | **0.26** | **0** |
| | | Temp Deviation | 0.04 | 0.215 | **0.241** | **0** | **0.065** | **0.042** | **0.25** | **0** |
| | 2 | Sleep Duration | 0.012 | 0.714 | **-0.084** | **0.01** | 0.031 | 0.329 | **-0.083** | **0.011** |
| | | REM Sleep | 0.037 | 0.252 | -0.013 | 0.69 | 0.031 | 0.327 | -0.048 | 0.146 |
| | | Deep Sleep | -0.024 | 0.462 | -0.02 | 0.533 | -0.018 | 0.584 | -0.031 | 0.35 |
| | | HRV (RMSSD) | -0.001 | 0.98 | -0.031 | 0.337 | 0.008 | 0.799 | **-0.051** | **0.12** |
| | | HR | **0.068** | **0.034** | **0.068** | **0.038** | 0.024 | 0.457 | **0.005** | **0.871** |
| | | RR | -0.051 | 0.111 | -0.008 | **0.816** | **0.054** | **0.092** | **0.117** | **0** |
| | | Temp Deviation | 0.013 | 0.693 | **0.068** | **0.039** | **0.021** | **0.507** | **0.076** | **0.02** |

**Device-Generated Metrics Adjusted for Pre-Vaccination Baseline Period**

Adjusting for the pre-vaccination baseline period strengthened associations between device-generated metrics and RBD antibody responses revealed significant associations between RBD antibody values and additional metrics (Table 2.4). Specifically, greater increases in HR and temp deviation on night 0 after the second injection adjusted for the pre-vaccination baseline period for both Moderna-NIAID (HR: rho = 0.148, p = 0.007; temp deviation: rho = 0.158, p = 0.004) and Pfizer-BioNTech (HR: rho = 0.124, p = 0.002; temp deviation: rho = 0.152, $p < 0.001$) were associated with greater RBD antibody responses. Additionally, on night 0 after the second injection, larger decreases in HRV and deep sleep, and a larger increase in RR, were associated with greater RBD antibody responses for Moderna-NIAID (HRV: rho = -0.119, p = 0.031; RR: rho = 0.139, p = 0.012) and Pfizer-BioNTech (HRV: rho = -0.182, $p < 0.001$; RR: rho = 0.114, p = 0.004). Pfizer-BioNTech also demonstrated an additional association between larger decreases in deep sleep and greater RBD antibody responses (Deep: rho = -0.120, p = 0.003). The associations between RBD antibody responses and both RR and temp deviation for each Pfizer-BioNTech and Moderna-NAIAD were also evident the following night (night 1) after the second injection. We did not observe these patterns for Johnson & Johnson-Janssen.

When we combined participants who received either of the two mRNA vaccines (Moderna-NIAID and Pfizer-BioNTech), the associations between device-generated metrics adjusted for pre-vaccination baseline and RBD antibody responses demonstrated greater statistical significance (Table 3). HR, HRV, RR, temp deviation, and deep sleep from night 0 after the second injection were significantly associated with RBD antibody responses. The associations between RBD antibody responses and HRV, HR, RR, and temp deviation were also statistically significant the following night (night 1) after the second injection. Additionally, greater HRV the night prior to the second injection (night -1) was significantly associated with greater RBD antibody responses (rho = 0.07, p = 0.024).

**Table 2.4.** Spearman rank order correlations between RBD antibody responses and device-generated metrics on nights before and after vaccine injections, adjusted for the pre-vaccination baseline period.

| Night Relative to Injection | Metric | Injection 1 J&J | | Injection 1 Moderna | | Injection 1 Pfizer | | Injection 2 Moderna | | Injection 2 Pfizer | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **rho** | *P* | **rho** | *P* | **rho** | *P* | **rho** | *P* | **rho** | *P* |
| -2 | Sleep Duration | -0.095 | 0.343 | -0.029 | 0.602 | 0.052 | 0.194 | **0.127** | **0.023** | 0.046 | 0.253 |
| | REM Sleep | -0.052 | 0.605 | -0.051 | 0.356 | **0.08** | **0.043** | 0.014 | 0.8 | 0.047 | 0.244 |
| | Deep Sleep | -0.137 | 0.173 | -0.011 | 0.841 | -0.008 | 0.847 | 0.073 | 0.194 | 0.028 | 0.478 |
| | HRV (RMSSD) | -0.105 | 0.297 | -0.035 | 0.526 | **-0.101** | **0.011** | 0.049 | 0.384 | 0.022 | 0.579 |
| | HR | -0.111 | 0.268 | 0.028 | 0.607 | **0.1** | **0.012** | -0.028 | 0.624 | -0.057 | 0.155 |
| | RR | -0.076 | 0.45 | -0.007 | 0.906 | 0.064 | 0.104 | **-0.047** | **0.402** | 0.037 | 0.357 |
| | Temp Deviation | -0.052 | 0.605 | 0.039 | 0.479 | 0.06 | 0.133 | 0.076 | 0.178 | -0.005 | 0.893 |
| -1 | Sleep Duration | 0.033 | 0.74 | -0.008 | 0.878 | -0.033 | 0.408 | -0.015 | 0.793 | -0.02 | 0.613 |
| | REM Sleep | -0.016 | 0.874 | -0.062 | 0.257 | -0.014 | 0.734 | -0.109 | 0.051 | -0.04 | 0.316 |
| | Deep Sleep | 0.062 | 0.539 | -0.029 | 0.594 | 0.07 | 0.079 | 0.011 | 0.847 | 0.018 | 0.659 |
| | HRV (RMSSD) | -0.077 | 0.441 | 0.035 | 0.522 | 0.004 | 0.92 | 0.054 | 0.339 | 0.049 | 0.22 |
| | HR | 0.128 | 0.199 | 0.01 | 0.853 | **-0.012** | **0.761** | -0.052 | 0.35 | **0.023** | **0.571** |
| | RR | 0.007 | 0.947 | -0.007 | 0.902 | **0.009** | **0.816** | 0.055 | 0.325 | **0.004** | **0.914** |
| | Temp Deviation | 0.004 | 0.971 | 0.031 | 0.567 | 0.003 | 0.944 | 0.072 | 0.201 | -0.017 | 0.678 |
| 0 | Sleep Duration | 0.166 | 0.098 | 0.097 | 0.079 | **-0.05** | **0.207** | 0.015 | 0.784 | 0.022 | 0.589 |
| | REM Sleep | 0.106 | 0.293 | 0.051 | 0.355 | -0.044 | 0.269 | 0.037 | 0.501 | 0.012 | 0.763 |
| | Deep Sleep | -0.173 | 0.083 | -0.01 | 0.863 | -0.009 | 0.816 | -0.054 | 0.326 | **-0.12** | **0.003** |
| | HRV (RMSSD) | -0.001 | 0.993 | -0.045 | 0.422 | 0.012 | 0.758 | **-0.119** | **0.031** | **-0.182** | **0** |
| | HR | 0.125 | 0.212 | -0.026 | 0.636 | **-0.018** | **0.642** | **0.148** | **0.007** | **0.124** | **0.002** |
| | RR | 0.113 | 0.262 | 0.008 | 0.881 | 0.074 | 0.061 | **0.139** | **0.012** | **0.114** | **0.004** |
| | Temp Deviation | 0.141 | 0.16 | 0.021 | 0.71 | 0 | 0.994 | **0.158** | **0.004** | **0.152** | **0** |
| 1 | Sleep Duration | **-0.224** | **0.025** | -0.068 | 0.215 | 0.026 | 0.508 | 0.088 | 0.12 | **0.081** | **0.045** |
| | REM Sleep | **-0.281** | **0.005** | 0.058 | 0.292 | -0.009 | 0.818 | **0.112** | **0.048** | **0.065** | **0.109** |
| | Deep Sleep | -0.152 | 0.13 | 0.024 | 0.667 | 0.047 | 0.238 | **-0.123** | **0.03** | 0.005 | 0.892 |
| | HRV (RMSSD) | 0.002 | 0.986 | -0.042 | 0.439 | -0.06 | 0.131 | -0.004 | 0.946 | -0.058 | 0.147 |
| | HR | 0.042 | 0.677 | -0.049 | 0.376 | **-0.005** | **0.896** | 0.042 | 0.46 | **0.054** | **0.182** |
| | RR | 0.057 | 0.573 | 0.103 | 0.061 | 0.041 | 0.299 | **0.168** | **0.003** | **0.259** | **0** |
| | Temp Deviation | 0.067 | 0.511 | 0.094 | 0.085 | -0.009 | 0.814 | **0.169** | **0.003** | **0.182** | **0** |
| 2 | Sleep Duration | 0.03 | 0.769 | -0.006 | 0.912 | 0.034 | 0.391 | -0.066 | 0.246 | **-0.064** | **0.111** |
| | REM Sleep | -0.125 | 0.214 | -0.018 | 0.74 | 0.016 | 0.685 | -0.08 | 0.155 | -0.031 | 0.446 |
| | Deep Sleep | 0.052 | 0.602 | -0.018 | 0.738 | -0.013 | 0.742 | 0.046 | 0.415 | -0.065 | 0.106 |
| | HRV (RMSSD) | -0.039 | 0.701 | 0.076 | 0.164 | -0.014 | 0.717 | 0.004 | 0.937 | **-0.112** | **0.005** |
| | HR | 0.111 | 0.269 | 0.01 | 0.858 | 0.021 | **0.602** | 0.003 | 0.963 | **0.033** | **0.408** |
| | RR | 0.082 | 0.416 | 0.099 | 0.069 | 0.028 | **0.477** | **0.138** | **0.014** | **0.097** | **0.016** |
| | Temp Deviation | 0.195 | 0.051 | 0.044 | 0.422 | -0.023 | 0.561 | **0.124** | **0.027** | 0.009 | 0.821 |

**Device-Generated Metrics during the Pre-Vaccination Baseline Period**

Among participants who received the Johnson & Johnson-Janssen vaccine, we did not observe any statistically significant ($p < 0.05$) correlations using Spearman rank order or Kendall rank correlation coefficients between baseline values of sleep duration, REM sleep, deep sleep, HRV, HR, RR, or temperature deviation and antibody responses. Among participants who received the Moderna-NIAID vaccine, we observed a correlation between temperature deviation and antibody response such that lower temperature deviation was associated with greater antibody response (Spearman rank order: rho = -0.108, p = 0.044; Kendall Tau: $\tau$ = -0.081, p = 0.049). Among participants who received the Pfizer-BioNTech vaccine, we observed a correlation between respiration rate and antibody response such that lower respiration rate was associated with greater antibody response (Spearman rank order: rho = -0.092, p = 0.017; Kendall Tau: $\tau$ = -0.064, p = 0.016).

**Multivariate models**

Based on the results of bivariate analyses, we focused multivariate analysis on night 0 after the second vaccine injection, using the combined mRNA vaccine (Pfizer-BioNTech and Moderna-NIAID) participants. In Cox regression models, temp deviation on night 0 after the second injection was a statistically significant predictor of RBD antibody responses in models before and after adjusting for the pre-vaccination baseline period (Table 5). In the model unadjusted for the pre-vaccination baseline period, greater HR on night 0 after the second injection was also a statistically significantly predictor of greater RBD antibody responses.

### 2.1.3 Discussion

We found that physiological metrics from an off-the-shelf wearable device on the two nights following the second dose of an mRNA-based COVID-19 vaccine were associated with RBD antibody responses. Using the device-generated metrics adjusted for the pre-vaccination baseline period, we found that both increased temperature deviation and heart rate (HR) and

**Table 2.5.** Multivariate regression models predicting RBD antibody responses from device-generated metrics before and after adjusting for the pre-vaccination baseline period that demonstrated associations with RBD antibody responses in Spearman correlations from night 0 after the second injection for Moderna-NIAID and Pfizer-BioNTech vaccine recipients, combined.

| **Device-Generated Metric**<br>**GFI –log2(p) = 44.95** | **Coefficient (Beta)** | **95% CI (LB, UB)** | **Z** | ***p*** |
|---|---|---|---|---|
| HRV (RMSSD) | -0.002 | (-0.007, 0.004) | -0.584 | 0.559 |
| RR | 0.035 | (-0.024, 0.093) | 1.161 | 0.246 |
| HR | -0.019 | (-0.031, -0.007) | -3.178 | 0.002 |
| Deep | 0 | (0.000, 0.000) | 1.125 | 0.26 |
| Temp Deviation | -0.515 | (-0.707, -0.322) | -5.247 | <.001 |
| **Device-Generated Metric Adjusted for Baseline period**<br>**GFI –log2(p) = 44.65** | | | | |
| HRV (RMSSD) | 0.047 | (-0.023, 0.117) | 1.308 | 0.191 |
| RR | -0.042 | (-0.091, 0.008) | -1.632 | 0.103 |
| HR | -0.013 | (-0.073, 0.047) | -0.428 | 0.669 |
| Deep | 0.036 | (-0.032, 0.104) | 1.043 | 0.297 |
| Temp Deviation | -0.071 | (-0.109, -0.034) | -3.748 | <.001 |

decreased heart rate variability (HRV) the night immediately following the second mRNA vaccine injection correlated with higher RBD antibody responses. In bivariate analyses using a standardized difference from the pre-vaccination baseline period for each physiological metric, we found that increased HR, temperature deviation, and RR, as well as decreased HRV and deep sleep, were each associated with higher RBD antibody responses for individuals who received the mRNA vaccines. We did not, however, find a meaningful pattern of associations between participants' device-generated metrics during the pre-vaccination baseline period and antibody responses. Although one earlier study did not find an association between vaccine-related side effects and antibody levels (Morales-Núñez et al., 2021), a second study did report higher antibody levels in individuals with clinically significant side effects (Debes et al., 2021). Neither study, however, reported actual antibody levels in relation to side effects. Importantly, rather than relying on participant-reported side effects, our study assessed objective (continuously assessed) physiological measures as predictors of vaccine responses. Data presented here speak to recent calls for examining wearable device data in tandem with immune responses to vaccines (Hajduczok et al., 2021). The continuous predictors in these analyses may have been more sensitive to the effects of vaccination in generating systemic inflammatory responses.

In a multivariate model predicting RBD antibody responses from device-generated metrics, we found that increased HR and temperature deviation independently predicted greater RBD antibody values for individuals who received the mRNA vaccines. In an identical model adjusting for the pre-vaccination baseline period, we found that dermal temperature was the sole statistically significant independent predictor of greater RBD antibody values in this same sample. Prior research has focused on identifying the roles of behavioral and psychological factors, in particular, sleep parameters, in vaccine responses. Short sleep duration prior to vaccination against influenza reduces antibody responses in both observational and experimental sleep deprivation models (Spiegel et al., 2002; Prather et al., 2021). These observations have driven hypotheses that a longer sleep duration prior to vaccination against COVID-19 might boost host immune responses (Benedict and Cedernaes, 2021). Ongoing research is studying the effects of shift work and short sleep on antibody response following mRNA-based COVID-19 vaccination (Lammers-van der Holst et al., 2022). Our data did not demonstrate associations between pre-vaccination sleep duration and antibody response among individuals receiving mRNA-based COVID-19 vaccines. In contrast to prior findings with other vaccines, we found that less deep sleep (NREM stages 3/4 sleep) the night immediately after receiving an mRNA-based vaccine against COVID-19, both in absolute terms and relative to one's pre-vaccination levels, was associated with greater antibody responses. Rather than implicating reduced deep sleep after vaccination as a mechanism driving antibody response, a more likely explanation is that individuals experiencing more noticeable discomfort, such as arm pain, fever, chills, or other symptoms that can follow COVID-19 vaccination (CDC, 2023) may have experienced sleep disruptions. Consistent with this explanation, sleep duration was not a significant predictor of vaccine responses in multivariate models. Future research should capture diversified self-reports of the effects of post-vaccination symptomology, including their perceived effects on sleep factors (e.g., duration, restfulness) to further explore this hypothesis.

Mediators of systemic inflammatory responses, such as COX-2, are associated with fever as well as with the development of certain vaccine responses (Saleh et al., 2016). As antipyretic

analgesics, including acetaminophen (Hinz et al., 2008) and non-steroidal anti-inflammatory drugs, can inhibit COX-2, these data have raised concerns that antipyretic analgesics might blunt certain vaccine responses if given at the time of vaccination. Randomized, controlled trial results in children have shown that prophylactic administration of antipyretic drugs at the time of vaccination led to significantly lower antibody responses to multiple vaccines (Prymula et al., 2009). Consistent with these data, the CDC recommends not taking antipyretic analgesics before the COVID-19 vaccination (CDC, 2024). It is also possible that the use of antipyretic analgesics following vaccination may blunt immune responses. However, few randomized controlled trials have formally examined this issue (Prymula et al., 2009). The CDC website suggests that one "talk to a doctor about taking over-the-counter medication, such as ibuprofen, acetaminophen, aspirin (only for people age 18 or older), or antihistamines for any pain and discomfort experienced after getting vaccinated" (CDC, 2023, 2024). To date, no trials that we are aware of have examined the role of fever or the impact of antipyretic medications on antibody responses following vaccination with mRNA vaccines. Although our data raise potential concerns that antipyretic analgesics might blunt COVID-19 vaccine responses, as temperature elevation was associated with greater antibody responses, our data did not directly address the effect of these medications and did not shed light on whether antipyretics influence immune pathways involved in the generation of immune responses to COVID-19 vaccines. Taken together, prior research (Prymula et al., 2009) and our data highlighted the potential importance of further research testing the effects of antipyretic medications used after receiving COVID-19 vaccines on antibody responses.

Psychological factors, including psychological stress, can impact immunological responses to vaccination, and clarifying their impact on COVID-19 vaccination may be important for developing interventions to optimize antibody response (Madison et al., 2021). For example, prior research demonstrated the negative effects of stress on immune response following vaccination against Hepatitis B (Glaser et al., 2 02). More recent work has shown negative effects of poor sleep prior to vaccination against influenza (Benedict and Cedernaes, 2021), and subsequent

studies have replicated similar patterns across multiple types of vaccinations (Madison et al., 2021). Decreased HRV can indicate increased sympathetic nervous system tone. Researchers have thus operationalized psychological stress using measures of heart rate variability (HRV), although the correlation of heart rate variability with stress is imperfect and can depend on several moderators, including contextual factors (Kim et al., 2018). We found that in the combined mRNA vaccine group, on the night immediately prior to the second injection adjusted for baseline (night -1), HRV was positively correlated with antibody responses. This suggests that lower HRV (consistent with greater psychological stress (Kim et al., 2018; Thayer et al., 2012)) was associated with lower antibody responses. This association changed direction in the following night (night 0) such that HRV was negatively correlated with antibody responses. This likely represents the effect of increased systemic inflammatory responses to vaccination, resulting in elevated temperature and HR, and decreased HRV, which in turn was associated with greater antibody responses. Consistent with this explanation, HRV did not significantly predict antibody response in multivariate models with other metrics (i.e., temperature), however, suggesting HRV after vaccination was not independently associated with antibody responses. Future research on this issue should measure stress more broadly (both self-report and physiological metrics of psychological stress) as predictors of antibody response to better clarify the role of stress in COVID-19 vaccine responses.

Our data have several limitations. We did not collect information on antipyretic or other medication use surrounding the time of vaccine injections, and thus cannot directly assess any effects of antipyretics on vaccine responses. We also did not collect detailed self-report information on post-vaccination symptomology, and we did not collect anthropometric information (e.g., height, weight, blood pressure) from participants as this study was completed by mail and internet only. Future research should collect such information, as emerging data suggests that health metrics, such as body mass index, may be associated with antibody responses (Uysal et al., 2022). The RBD antibody assay we used had an upper limit of dynamic range of 2500 IU/mL. A substantial proportion of participants achieved antibody levels above this range,

resulting in right-censored data. To address this, we used Cox regression, a robust approach for analyzing right-censored data (readers may be more familiar with this approach when analyzing time-to-event data). Future research would benefit from an antibody assay with an extended dynamic range. We used a commercially available RBD antibody assay to assess neutralizing antibody responses rather than more precise but more expensive approaches, such as pseudo virus neutralization assays. Like other studies exploring vaccination responses, our results will become more meaningful once there are enough data to form a scientific consensus on an antibody level that indicates adequate immune protection.

## 2.2 Physiological changes around fever onset

### 2.2.1 Materials and Methods

We previously reported on data collected for these analyses by Mason et al. (Mason et al., 2022). Additional details on the recruitment and exclusion criteria of the initial cohort are outlined in Mason et al.; however, we outline details relevant to the subset of participants used in these analyses. The original cohort comprised 63,153 participants spanning 106 countries (Mason et al., 2024) who completed online questionnaires and wore the Oura Ring Gen2, a commercially available wearable device (Oura Health, Oulu, Finland) on a finger of their choosing. Participants completed baseline, monthly, and daily online questionnaires; the daily questionnaire included a checklist to report the subjective experience of a number of symptoms. These analyses focused on self-reported fever symptoms; participants could self-report the symptom "Fever" since they last completed a daily questionnaire ("*Have you experienced any of the following symptoms since you last did this survey? (Please check all that apply.)*"). Participants were also asked to self-report the highest body temperature reading they had taken during the last day by thermometry (*"If you took your temperature in the last day, what was the highest reading?"*). To select days that were more likely to be from a fever event, we considered any day where a participant reported both (1) experiencing a self-reported fever and (2) a self-reported temperature greater than or equal to

38 °C to be a fever day. Fever days with wearable device data from at least seven nights over a fourteen-day baseline period and the nights before and after the fever day were included in the dataset. Wearable device data from the nights before and after fever days comprised positive class examples in the training set and the test set. Negative class examples comprised days wherein participants both (1) self-reported not experiencing fever and (2) self-reported a temperature lower than 38 °C (non-fever days). Non-fever days also had retrievable wearable device data from at least seven nights over a fourteen-day baseline period and the nights before and after the non-fever day. Participants wore the Oura Ring Gen2 (Oura Health Oy, Oulu, Finland). The Oura Ring connects to the Oura App (available from the Google Play Store and the Apple App Store) via Bluetooth. Users can wear the ring continuously in both wet and dry conditions. The Oura Ring generates physiological metrics by aggregating data gathered from on-device sensors. These high-resolution metrics are transformed into summary metrics before their transmission to a smartphone app. The Oura Ring Gen2 uses a proprietary algorithm to estimate when a user is at rest and when they have gone to bed. After the Oura Ring determines that a user has gone to bed, the Oura Ring gathers a high-frequency photoplethysmogram (PPG), which it uses to calculate interbeat intervals (IBI), which are used in heart rate (HR), heart rate variability (HRV), and respiratory rate (RR). Both HR and HRV measured by Oura have been externally validated to be highly accurate (Cao et al., 2022). RR has been validated internally by Oura and is claimed to be highly accurate compared to a medical-grade ECG, with a mean error of 0.71 breaths per minute and a correlation of 0.96 (Kryder, 2020b). The Oura Ring Gen2 assesses a user's dermal (distal) temperature throughout the day (i.e., not only when the user is in bed) using a negative temperature coefficient (NTC) thermistor on the internal surface of the ring. The NTC thermistor has been internally validated by Oura and has been shown to provide near-perfect agreement with a research-grade sensor (Kryder, 2020a). During sleep, the Oura Ring uses a proprietary algorithm to estimate the stage of sleep a user is currently in. Sleep stages can be one of the following: awake, REM, light (N1 or N2), or deep (N3). This algorithm has been externally validated and is 79% accurate for four-stage sleep stage classification (Altini and Kinnunen,

2021). Further details regarding these sensors and the algorithms used to determine HR, HRV, RR, and sleep stages are provided in the Introduction. High-resolution metrics are transformed into summary metrics before transmission to a smartphone app. Oura further aggregates these summary metrics across each period of detected sleep into a "sleep summary". The dataset used in these analyses comprises metrics ("sleep summary metrics") from the longest sleep of the day (i.e., the sleep summary with the greatest total time spent asleep). We included all sleep summary metrics generated by Oura that were single, scalar, and physiologically interpretable values. Sleep summaries also included metrics that we did not include, i.e., arrays of HR and HRV across every 5 min of sleep, strings that specify the start and end of detected bedtimes, or any of the metrics that are a proprietary combination of the metrics we included (i.e., so-called "sleep scores"). Table 2.6 lists each sleep summary metric included in these analyses, along with detailed descriptions.

### 2.2.2 Results

Sixteen thousand, seven hundred, and ninety-four participants provided at least one valid ground truth day; there were a total of 724 fever days (positive class examples) from 463 participants and 342,430 non-fever days (negative class examples) from 16,687 participants. The mean self-reported body temperature was 38.45 (SD = 0.50) for fever days and 36.45 (SD = 0.42) for non-fever days. The distributions of self-reported body temperatures can be found in Figure 2.3. Table 2.7 provides the characteristics of participants included in these analyses. The average participant age was 47.2 years; 43.6% were women.

Wearable-measured physiological changes in the nights before and after fever days appear in Figure 3. Relative to individuals' wearable-measured baseline physiology, wearable-measured physiology changed substantially on the nights before and after self-reported fever days (Figure 2.4) and exhibited greater deviations in the subset of participants (n = 103) with fever days in which self-reported temperatures were greater than 39 °C (red lines, Figure 2.4). Across all participants with fever days, wearable measured physiology changed the most on the

**Table 2.6.** Detailed descriptions of each wearable measured sleep summary feature.

| Metric | Unit of Measurement | Description |
|---|---|---|
| Heart rate | Beats per minute | The average heart rate registered during the sleep period. |
| Lowest heart rate | Beats per minute | The lowest heart rate (5 min sliding average) registered during the sleep period. |
| Heart rate variability | Milliseconds | The average HRV calculated using the rMSSD method. |
| Respiratory rate | Breaths per minute | Average respiratory rate. |
| Respiratory rate variability | Breaths per minute | The average variability of respiratory rate (STD) in the sleep period. |
| Temperature deviation | Degrees Celsius | Skin temperature deviation from the user's long-term temperature average. |
| Temperature trend deviation | Degrees Celsius | Skin temperature deviation from weighted three-day rolling temperature average. |
| Onset latency | Seconds | Detected latency from the time the user entered their bed to the beginning of the first five minutes of persistent sleep. |
| Time spent awake | Seconds | Total amount of awake time registered during the sleep period. |
| Time spent in REM sleep | Seconds | Total amount of REM sleep registered during the sleep period. |
| Time spent in light sleep | Seconds | Total amount of light (N1 or N2) sleep registered during the sleep period. |
| Time spent in deep sleep | Seconds | Total amount of deep (N3) sleep registered during the sleep period. |
| Time spent asleep | Seconds | Total amount of sleep registered during the sleep period. |

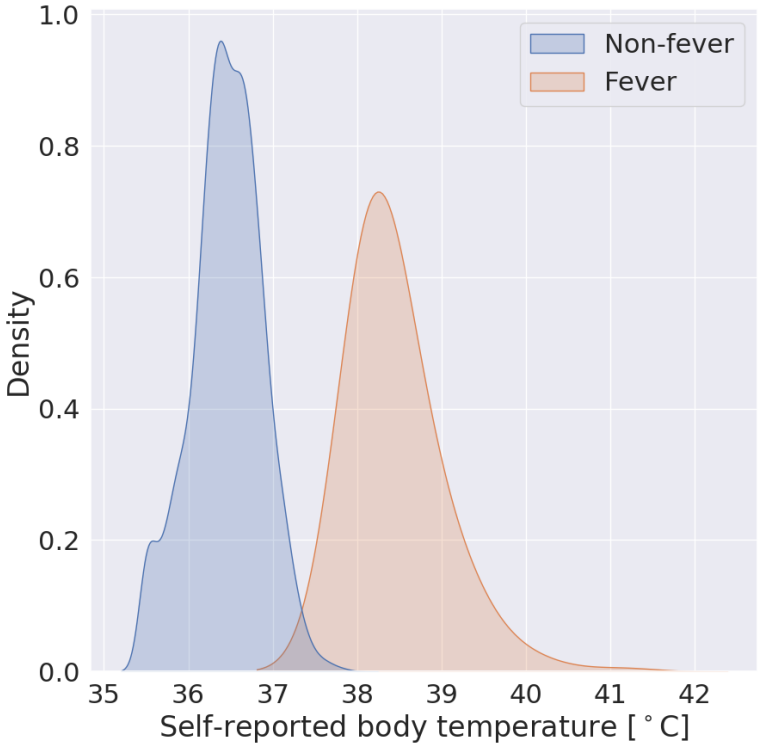


**Figure 2.3.** Self-reported body temperatures from non-fever examples are in blue and fever examples are in orange.

**Table 2.7.** The number of individuals included in the training and test sets, including self-reported sex assigned at birth, age, and race.

| | | **Dataset composition** |
|---|---|---|
| N | | 16,794 |
| Sex, n (%) | Female | 7,324 (43.6) |
| | Male | 9,455 (56.3) |
| | Other | 15 (0.1) |
| Age, mean (SD) | | 47.2 (12.3) |
| Race, n (%) | African American/Black | 226 (1.4) |
| | East Asian | 685 (4.2) |
| | Caucasian/White | 14,120 (86.3) |
| | Middle Eastern | 94 (0.6) |
| | Native American/Native Alaskan | 27 (0.2) |
| | Native Hawaiian or Other Pacific Islander | 28 (0.2) |
| | South Asian | 162 (1.0) |
| | Other | 429 (2.6) |
| | Prefer not to answer | 596 (3.6) |

nights before and after fever days (Nights -1 and 0, Figure 2.4).

## 2.3 Conclusion

Here, we demonstrate the feasibility of characterizing two distinct acute physiological changes using wearable device data. In both vaccination and fever onset, the maginitude of physiological changes are correlated with clinically important outcomes (antibody production and fever body temperature, respectively). In Chapter 2, I develop and characterize a machine learning model for detecting fever onset using these data.

## 2.4 Acknowledgments

Chapter 2, contains information as it appears in A.E. Mason, P. Kasl, W. Hartogensis, J.L. Natale, S. Dilchert, S. Dasgupta, S. Purawat, A. Chowdhary, C. Anglo, D. Veasna, L.S. Pandya, L.M. Fox, K.Y. Puldon, J.G. Prather, A. Gupta, I. Altintas, B.L. Smarr, and F.M. Hecht, "Metrics from Wearable Devices as Candidate Predictors of Antibody Response Following Vaccination against COVID-19: Data from the Second TemPredict Study", 2022 *Vaccines*. The dissertation author was a co-first author on this paper. It also contains information as it appears in P. Kasl, L. Keeler Bruce, W. Hartogensis, S. Dasgupta, L.S. Pandya, S. Dilchert, F.M. Hecht, A. Gupta, I. Altintas, A.E. Mason, and B.L. Smarr, "Utilizing wearable device data for syndromic surveillance: A fever detection approach", 2024 *Sensors*. The dissertation author was the primary author of this paper.

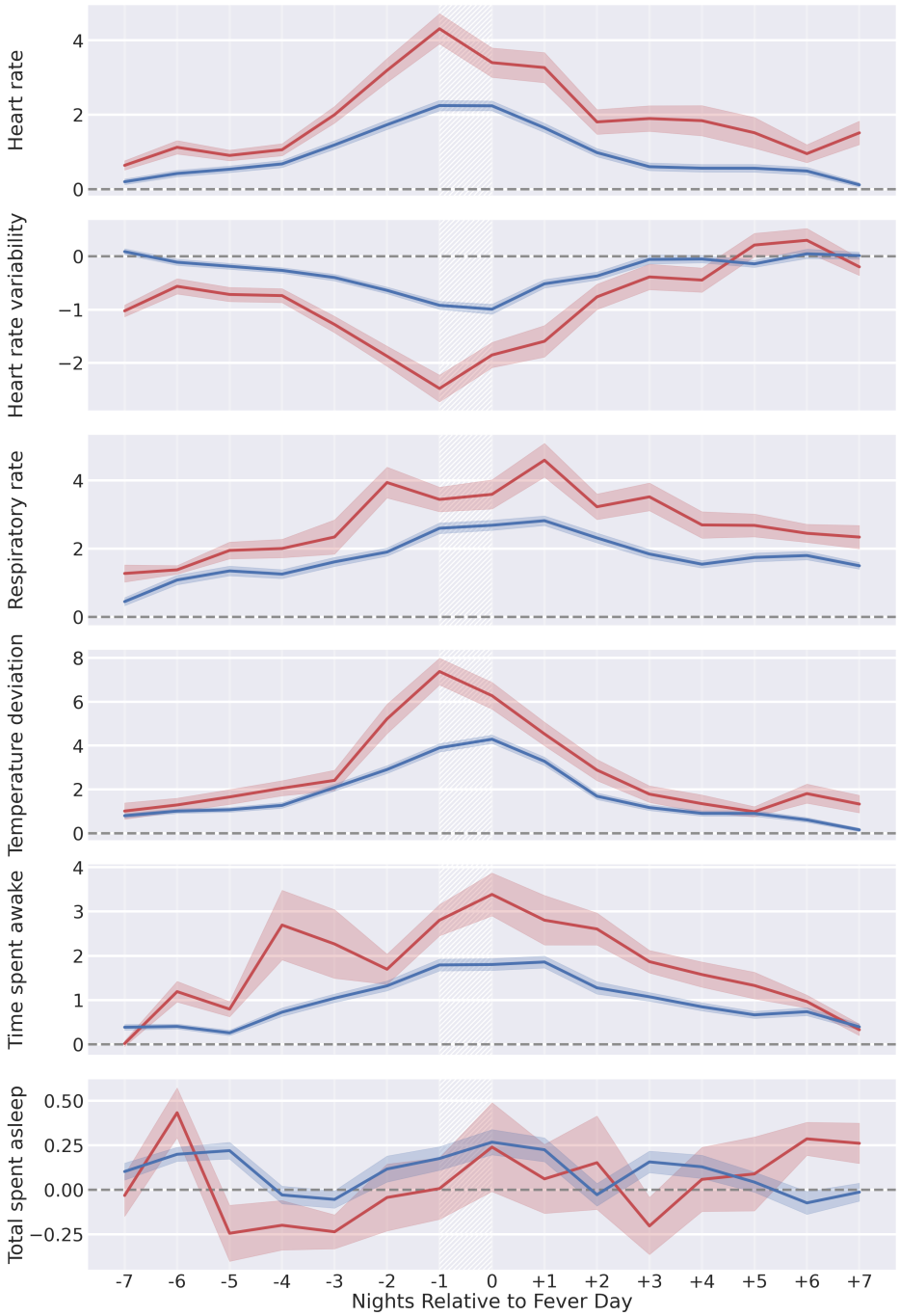


**Figure 2.4.** Z-score-normalized wearable metrics from individuals, aligned by self-reported fever day (white hatched areas) and grouped by self-reported temperature on fever day. Individuals reporting temperatures in the range of (38–39 °C) are in blue (n = 621), and (39+ °C) are in red (n = 103). Lines represent the mean z-score normalized wearable metric across all participants in the respective group for each night, and shaded regions are the 95% confidence interval of the mean.

# Chapter 3

# Developing a model for fever onset detection using wearable data

Commercially available wearable devices (wearables) show promise for continuous physiological monitoring. Previous works have demonstrated that wearables can be used to detect the onset of acute infectious diseases, particularly those characterized by fever. We aimed to evaluate whether these devices could be used for the more general task of syndromic surveillance. We obtained wearable device data (Oura Ring) from 63,153 participants. We constructed a dataset using participants' wearable device data and participants' responses to daily online questionnaires. We included days from the participants if they (1) completed the questionnaire, (2) reported not experiencing fever and reported a self-collected body temperature below 38 °C (negative class), or reported experiencing fever and reported a self-collected body temperature at or above 38 °C (positive class), and (3) wore the wearable device the nights before and after that day. We used wearable device data (i.e., skin temperature, heart rate, and sleep) from the nights before and after participants' fever day to train a tree-based classifier to detect self-reported fevers. We evaluated the performance of our model using a five-fold cross-validation scheme. Sixteen thousand, seven hundred, and ninety-four participants provided at least one valid ground truth day; there were a total of 724 fever days (positive class examples) from 463 participants and 342,430 non-fever days (negative class examples) from 16,687 participants. Our model exhibited an area under the receiver operating characteristic curve (AUROC) of 0.85 and

an average precision (AP) of 0.25. At a sensitivity of 0.50, our calibrated model had a false positive rate of 0.8%. Our results suggest that it might be possible to leverage data from these devices at a public health level for live fever surveillance. Implementing these models could increase our ability to detect disease prevalence and spread in real-time during infectious disease outbreaks.

## 3.1 Introduction

Public health agencies commonly use syndromic surveillance (SS) to augment a variety of traditional disease surveillance systems (Mandl et al., 2004; Smith et al., 2019). SS systems generally do not assess laboratory-confirmed reports and instead rely on the presence of detectable symptoms; cases are typically reported before the results of a laboratory test are available (Mandl et al., 2004). SS systems require a lower implementation burden relative to traditional surveillance systems that rely on case reports, such as the National Notifiable Disease Surveillance System. SS systems are, therefore, potentially (1) more scalable, (2) more sensitive, and (3) better able to more rapidly identify outbreaks (Colón-González et al., 2018; Henning, 2004). Systems using commercially available wearable devices (wearables) to detect illness states exhibit many of the same strengths as SS. That is, they are (1) scalable, as in 2019, approximately 30% of US consumers already used wearables, which are relatively inexpensive (Chandrasekaran et al., 2020); (2) sensitive as wearable device physiological data can be monitored in large, distributed, diverse populations, and can be used to discern periods of relative health versus illness; and (3) rapid as wearable device data can be analyzed in near real-time. Many recent efforts propose machine learning classifiers for the within-individual detection of specific, acute illnesses using wearable device data (Mason et al., 2022; Alavi et al., 2022; Gadaleta et al., 2021; Richards et al., 2021; Miller et al., 2020; Grzesiak et al., 2021; Chaudhury et al., 2022; Mitratza et al., 2022; Smarr et al., 2020; Merrill et al., 2023). Other works have investigated using wearables to monitor population-level changes corresponding to influenza-like illnesses (ILI) (Konty et al.,

2019; Mezlini et al., 2022). Both within-individual detection and population-level monitoring tasks are tractable because wearables measure physiological metrics that are anomalous around acute illness onset. These anomalies can include increased heart rate (HR), respiratory rate (RR), and temperature, and decreased heart rate variability (HRV) and physical activity (Mitratza et al., 2022). However, real-time SS systems hold the potential to detect such aberrations that may signal the increased prevalence of a novel pathogen (Smith et al., 2019). As such, we sought to determine whether wearable device data could be used for generalized SS, and we evaluated such feasibility by focusing on fever detection. Fever is often a crucial component of the case definition for many SS systems across conditions, including ILI, where the presence of fever is necessary but not sufficient for a case to be considered an ILI event (Fitzner et al., 2018). Moreover, fever is sometimes the only symptom surveilled (Shimoni et al., 2012, 2008; Hiller et al., 2013). In this work, we explored changes in wearable-measured physiology around the onset of self-reported fevers, proposed a classifier for detecting its onset, and demonstrated the classifier's performance in a broad population.

## 3.2 Methods

The input features to our model follow the standard format for a binary classification task. Let $D = \{(x_1, y_1) \ldots (x_n, y_n)\}$ be the training dataset. $x_j \in R_k$ and $y_j \in \{0, 1\}$. $x_j$ is a vector of size $k = 35$. Entries 1,..., 14 in $x_j \stackrel{\text{def}}{=} z_{i,m}$ are as follows

$$z_{i,m} = \frac{Night_{i,m} - \mu_{(-14 \rightarrow -28),m}}{\sigma_{(-14 \rightarrow -28),m}} \tag{3.1}$$

Here, the z-scored wearable device metrics from the night before (Night -1, Figure 3.1) are from the ground truth day. Similarly, entries 15,..., 28 in $x_j \stackrel{\text{def}}{=} z_{i,m}$ are from the night after (Night 0, Figure 3.1) the ground truth day. Entries 29,..., 35 in $x_j \stackrel{\text{def}}{=} \in \{0, 1\}$ correspond to one-hot-encoded Boolean features for the day of the week (Sunday through Monday) of the ground truth day. In summary, the features are (1) z-scored sleep summary metrics ($x_{i,m}$) from the

night before (NB) and the night after (NA) each fever or non-fever day and (2) one-hot-encoded Boolean features for the day of the week (Sunday through Monday) of the ground truth day. We included the day of the week as a feature, given the tendency for human weekly rhythms (i.e., alcohol consumption (Alavi et al., 2022)) to drive physiological changes that manifest similarly to acute illnesses. $y_j = 0$ if the jth example is from a non-fever day and $y_j = 1$ if the jth example is from a fever day. A schematic describing the normalization procedure and instance selection process is shown in Figure 3.1.

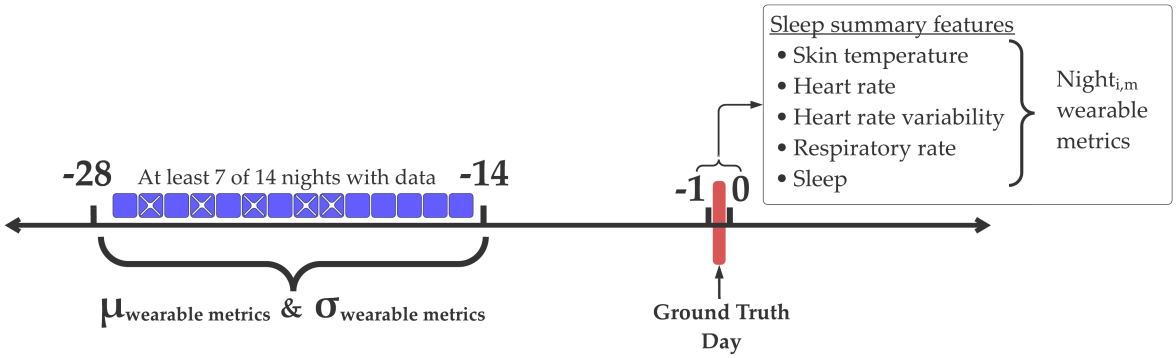


**Figure 3.1.** Instance selection and normalization procedure. At least 7 out of the 14 days in the range of -28 to -14 relative to the ground truth day were retrievable. The mean ($\mu$) and standard deviation ($\sigma$) from these days were used to normalize z-score wearable device metrics. We depict an example of a valid instance with its baseline period (-28 $\rightarrow$ -14) with retrievable data from 9 out of 14 nights (nights without retrievable data are indicated by a white cross). This instance is based on sleep summary features from the night before (night -1) and the night after (night 0) relative to the ground truth day.

In order to ensure applicability, we implemented a relatively simple, commonly used ensemble classifier based on the standard implementation of a Histogram-Based Gradient-Boosting Classification Tree from the sklearn Python (Open source) package v1.2.0 (sklearn.ensemble.HistGradientBoostingClassifier) with all hyper-parameters left at default. Models of this variety are commonly used for physiological anomaly detection (Gadaleta et al., 2021; Miller et al., 2020; Nestor et al., 2023; Conroy et al., 2022). For training and testing, we followed a five-fold stratified cross-validation scheme with a user split as previously outlined in Merill et al. (Merrill et al., 2023), where each model was trained on data from a subset of participants and tested on

another subset. We stratified users based on whether that user had a fever day.

Classifiers can be calibrated during training, which aligns a classifier's predicted class probabilities and the empirical likelihood of events occurring (Vaicenavicius et al., 2019). Predictions from well-calibrated classifiers tend to more accurately reflect real-world outcomes. Importantly, this can allow practitioners to choose intervention thresholds based on a classifier's predictions, which can lead to more precise resource allocation and risk assessment (Sahoo et al., 2021). We used logistic (sigmoid) regression with a two-fold split to calibrate our model using the sklearn v1.2.0 implementation (sklearn.calibration.CalibratedClassifierCV). We used the Brier score to assess the extent to which our classifier was calibrated (Roulston, 2007). The Brier score was calculated by taking the squared difference between the classifier's predicted probability and the corresponding outcome (0 for incorrect predictions and 1 for correct ones). The Brier score was then the mean squared difference across all predictions. Brier scores ranging from 0 to 1 and lower values indicate a more calibrated classifier. We used the sklearn v1.2.0 implementation of the Brier score (sklearn.metrics.brier_score_loss).

We examined the relative importance of each wearable and measured physiological change in our classifier using permutation importance, which is a data-driven approach that quantifies the weight that a tree-based classifier places on individual features (Breiman, 2001). Permutation importance is determined by evaluating how much a classifier's performance degrades after the systematic perturbation of a specific feature. Baseline classification performance is established on the unperturbed dataset. Then, each individual feature (i.e., the z-score and average HR from the night before a [non]-fever day) is randomly permuted between examples (i.e., all [non]-fever days) in the dataset. This permutation disrupts any relationship between the feature and the classification output. The change in classification performance is determined after permutation. Features, when permuted, that cause the largest drop in classification performance are the most important. We used the sklearn v1.2.0 permutation importance (sklearn.inspection.permutation_importance) with 30 permutations per feature at each iteration of the five cross-validation.

The receiver operating characteristic (ROC) and Precision-Recall curves are often used to visually assess binary classification performance (Su et al., 2015). The ROC illustrates the relationship between a classifier's true positive rate (i.e., recall, sensitivity) and false positive rate (i.e., 1-specificity) across predicted probability threshold values. The ROC curve is often used to examine the trade-off between correctly identifying positive instances and incorrectly classifying negative instances as positive. The integration of the ROC yields the area under the ROC (AUROC), which is commonly used to summarize the ROC. On the other hand, the Precision-Recall curve (PRC) plots precision (i.e., positive predictive value) against recall (i.e., true positive rate, sensitivity) across predicted probability threshold values. The PRC can more accurately represent the performance on imbalanced datasets; this method describes a classifier's ability to correctly identify positive examples while minimizing false positives. Average (i.e., mean) precision (AP) is frequently used to summarize the PRC.

## 3.3 Results

We depicted model performance following a five-fold cross-validation scheme in Figure 3.2. The mean AUROC was 0.85 (Figure 3.2), and the mean AP was 0.25 (Figure 3.2). Our model was well calibrated (Figure 3.2) with a Brier score of 0.0018. When considering the aggregated predictions on the test set of each cross-validation, the positive class predicted that probabilities increased with increased self-reported body temperature (Figure 3.2) and were significantly correlated (Pearson's r = 0.11, $p < 0.001$); at a sensitivity of 0.50, the false positive rate was 0.8%.

We calculated the permutation importance at each iteration of the five cross-validations. Permutation importance suggested that temperature deviation from the night before a fever day was the most important feature (Figure 3.3), followed by respiratory rate and the time spent awake the night before the ground truth day.

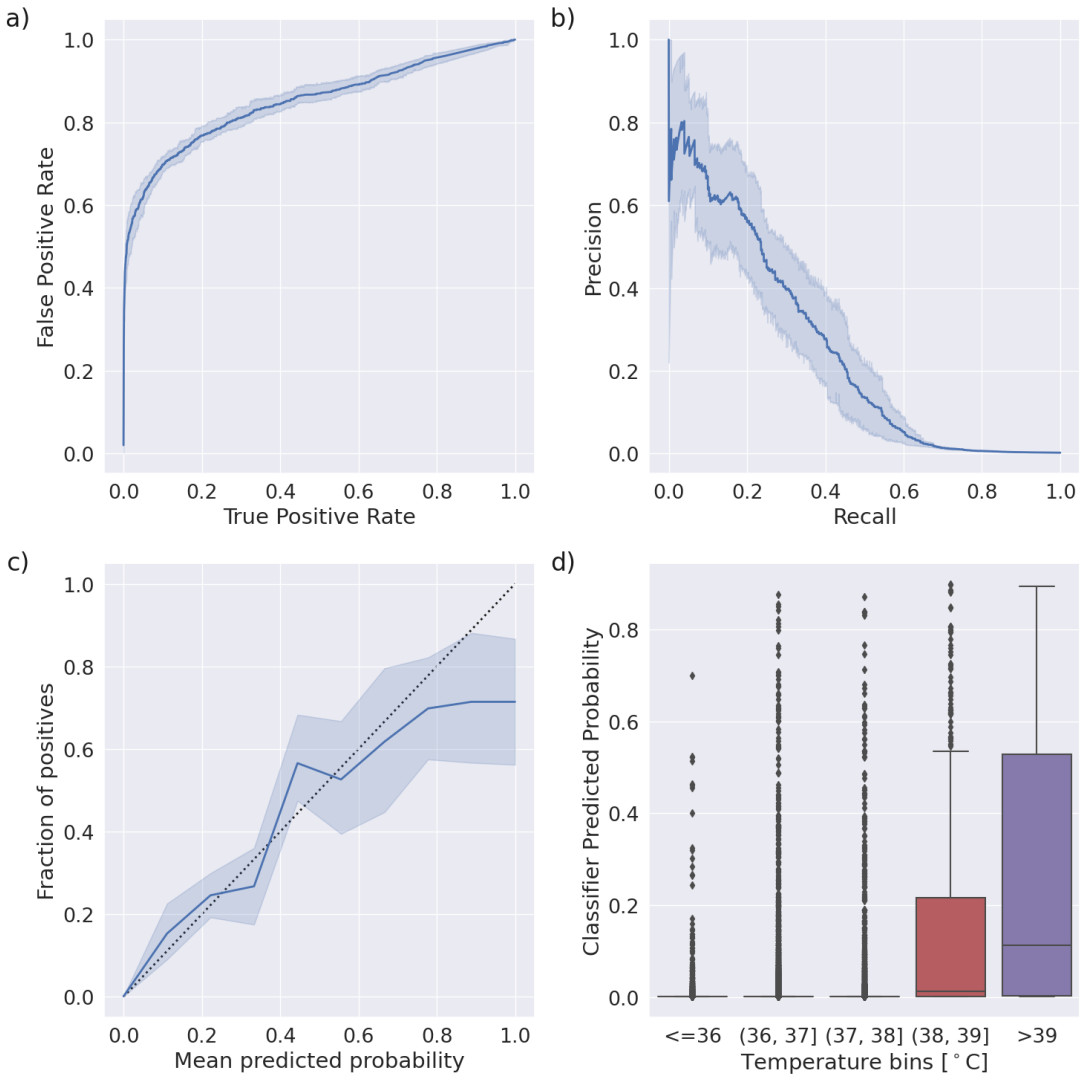


**Figure 3.2.** Performance of the fever detection classifier following a five-fold cross-validation scheme. Shaded areas indicate a 95% confidence interval. (a) The mean Receiver Operator Characteristic curve (ROC) across iterations. The mean area under the curve is 0.85. (b) The mean Precision–Recall curve (PRC) across iterations. The average precision was 0.25. (c) The reliability plot (or calibration curve) across iterations. The mean Brier score was 0.0018. (d) Box plots indicating the classifier predicted probability, binned by self-reported body temperature.

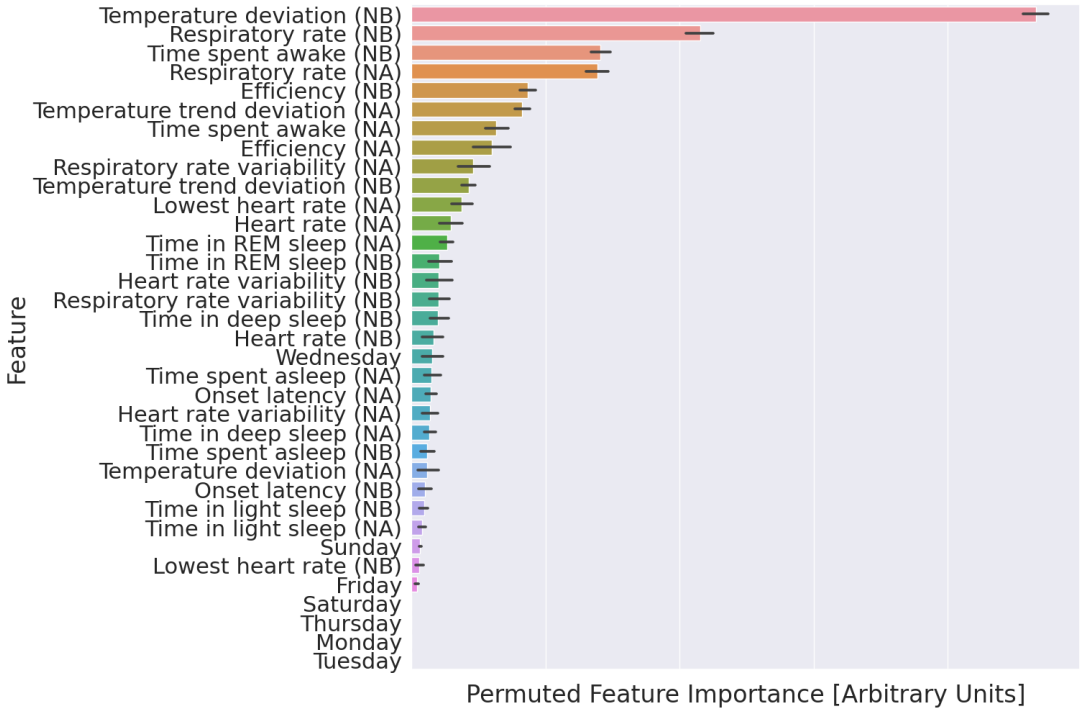


**Figure 3.3.** Explanation of the fever detection classifier. Features are ranked from most (top) to least (bottom) important based on the mean permuted importance across 30 permutations. NB: Night before [non]-fever day; NA: night after [non]-fever day; days of the week (i.e., Sunday) indicate the ground truth day; error bars: 95% confidence interval of the mean.

## 3.4 Discussion

We found support for the hypothesis that data from wearable devices can be used to detect fevers with high accuracy on the night after the day an individual starts to experience a fever. Specifically, we described wearable measured physiological changes around fever onset (Figure 3) and developed features that were quite computationally tractable and had direct physiological interpretations. Our classifier performed well (average AUROC = 0.85, AP = 0.25) and could be tuned to a sensitivity of 0.50, where it exhibited a false positive rate of 0.8%.

Over a large population, detection using wearable devices could provide important new alerting functionality to SS efforts. Since our model inclusion criteria only required retrievable wearable device data over a two-week baseline period, our model could make predictions on any new device users after about a month of continuous wear time. We calibrated our classifier so that higher predicted examples were more likely to be from a fever day, and our classifier could show promise for a body temperature regression task; the predicted probability increased proportionately to the self-reported body temperature that described a fever. We posit that features with explicit physiological interpretations allow better generalizability to heterogeneous populations than features learned by deep neural networks using a similarly sized training set and believe this to be a key next step following from this work.

Readers should interpret these results in light of our classifier implementation, performance metrics selection, and definition of illness and non-illness periods. While our classifier exhibited sensitive and specific fever onset detection using wearable-measured physiological data in a diverse population, further testing should systematically compare the current classifier implementations across a range of benchmark datasets to determine which classifiers should be further evaluated for deployment. We chose a machine learning architecture that was relatively simple and common to train our classifier; however, there is a wide diversity of approaches used to classify illness from wearable device data (for review, see Mitratza and colleagues) (Mitratza et al., 2022). Furthermore, certain binary classification performance metrics (i.e., AUROC,

accuracy) can lead to misleading notions of performance when used on datasets that exhibit extreme class imbalance, as in these analyses where the number of non-fever days far outnumber fever days. Such a class imbalance is common in illness detection studies (Nestor et al., 2023). Accordingly, we attempted to report all metrics in a way that did not overestimate the performance. A systematic comparison of illness detection classifiers would require consistent definitions of illness and non-illness periods across benchmark datasets, as well as the use of the same metrics to describe classifier performance across these datasets.

This work also differs from other illness detection studies in both study design and the wearable device used to gather data. We performed these analyses retrospectively, and the performance should be verified in a prospective manner (Nestor et al., 2023). Furthermore, differences in commercially available wearable device sensors (i.e., the ability to collect HRV, HR, temperature, and other physiological metrics) have led to substantial differences in the features used to train illness detection classifiers. We trained our classifier using data from second-generation Oura Rings, which, at the time of data collection, were different from most other wearable devices in that they included a temperature sensor, which was not included in most other wearable devices of similar cost and market penetration (i.e., Apple Watch and FitBit). Regardless of feature differences, data from wearable devices without temperature sensors have been used to train many of the other previously studied illness detection classifiers over the past decade (Alavi et al., 2022; Merrill et al., 2023). However, many of the most recent generations of wearable devices from Apple, FitBit, and Whoop now include a temperature sensor. Future work should investigate if and how different sensors in wearable devices create features that improve illness detection performance, particularly because our results suggest that temperature sensor-based features are the most important in our classifier (Figure 5). Measurements from sensors not traditionally included in commercial wearable devices, such as those that monitor analytes in sweat (Jagannath et al., 2021) or exhaled air (Shih et al., 2010), might be particularly important for improving the accuracy of these models. Other efforts have engineered more complicated features, i.e., features based on deviations from expected circadian rhythms (Hirten

et al., 2021); here, we demonstrate an impressive performance using nightly summary data. Researchers should systematically explore the effects of the study design and wearable device features as they work toward developing standards of real-world efficacy.

Our specific algorithmic implementation requires a minimum level of wearable device compliance. Previous work based on the dataset we used here demonstrates that participants exhibit a high level of wearable device compliance (87.8% of nights) (Shiba et al., 2023). Another survey-based study found that 72.58% of participants in their study wore their wearable device "daily" or "almost daily" (Chandrasekaran et al., 2020). Future work could weigh certain metrics like recall against the proportion of days wherein users provide enough data to produce variable results in order to determine the efficacy of these models.

As with other health-screening applications, illness detection algorithms based on wearable device data need to balance improving case detection with minimizing false positives. Illness detection generalizability should also be carefully evaluated across classifier implementations, the wearable devices used, and diverse populations. In particular, researchers should address whether models generalize across geographic regions. Future work should also examine whether the performance of illness detection models varies temporally. Such temporal performance variability might be driven by seasonality in illness prevalences. Once models exhibit a performance that can have a real-world impact, developments in wearable device data deidentification and data integration at public health agencies will be crucial to developing systems for real-time illness monitoring. Data privacy and deidentification are challenges that remain largely unaddressed for wearable device data. Recent works further demonstrate how it might be possible to re-identify individuals using de-identified wearable device data (Chikwetu et al., 2023). Furthermore, as of 2024, these data fall under the category of "personal health data" in the EU (EUROPEAN PARLIAMENT, 2016) and US (104th Congress, 1996), and these data are subject to regulations that vary by jurisdiction. However, it is possible that the categorization of these data might change in the future, along with the regulations they are subject to. Finally, our efforts suggest that symptom screening classifiers that generalize across illnesses may be a useful public health

tool for real-time surveillance.

## 3.5 Acknowledgments

Chapter 3, contains information as it appears in P. Kasl, L. Keeler Bruce, W. Hartogensis, S. Dasgupta, L.S. Pandya, S. Dilchert, F.M. Hecht, A. Gupta, I. Altintas, A.E. Mason, and B.L. Smarr, “Utilizing wearable device data for syndromic surveillance: A fever detection approach”, 2024 *Sensors*. The dissertation author was the primary author of this paper.

# Chapter 4

# A cross-study analysis of wearable datasets and the generalizability of acute illness monitoring models

Large-scale wearable datasets are increasingly being used for biomedical research and to develop machine learning (ML) models for longitudinal health monitoring applications. However, it is largely unknown whether biases in these datasets lead to findings that do not generalize. Here, we present the first comparison of the data underlying multiple longitudinal, wearable-device-based datasets. We examine participant-level resting heart rate (HR) from four studies, each with thousands of wearable device users. We demonstrate that multiple regression, a community standard statistical approach, leads to conflicting conclusions about important demographic variables (age vs resting HR) and significant intra- and inter-dataset differences in HR. We then directly test the cross-dataset generalizability of a commonly used ML model trained for three existing day-level monitoring tasks: prediction of testing positive for a respiratory virus, flu symptoms, and fever symptoms. Regardless of task, most models showed relative performance loss on external datasets; most of this performance change can be attributed to concept shift between datasets. These findings suggest that research using large-scale, pre-existing wearable datasets might face bias and generalizability challenges similar to research in more established biomedical and ML disciplines. We hope that the findings from this study will encourage discussion in the wearable-ML community around standards that anticipate and account for

challenges in dataset bias and model generalizability.

## 4.1 Introduction

Commercially available wearable devices (wearables) offer a unique, real-world, highly temporally resolved lens into an individual's physiology across time. Wearables continuously monitor several physiological signs (e.g., heart rate, step counts, sleep). Researchers increasingly view these signs as informative of an individual's health status. Several cross-sectional observational studies have correlated these signs with certain human conditions (e.g., step counts with incident disease (Master et al., 2022) and sleep with psychiatric conditions (Wainberg et al., 2021)). Measuring signs with wearables might also enable real-time health monitoring and even early intervention if machine learning (ML) models can predict health status changes before individuals become aware of them. Accordingly, numerous studies have demonstrated substantial progress towards using ML models trained on wearable data for a variety of within-individual longitudinal monitoring tasks including mental health conditions (e.g., depression (Xu et al., 2022a), anxiety (Wainberg et al., 2021)), chronic diseases (e.g., diabetes (Lam et al., 2021), sleep apnea (Master et al., 2022)), and specific acute illnesses (e.g., COVID-19 (Goergen et al., 2022; Abir et al., 2022; Richards et al., 2021; Gadaleta et al., 2021; Conroy et al., 2022; Yamagami et al., 2021; Hirten et al., 2021; Natarajan et al., 2020; Mayer et al., 2022; Alavi et al., 2022; Miller et al., 2020; Pho et al., 2023; Hirten et al., 2022), influenza (flu; Merrill et al., 2023; Grzesiak et al., 2021; Mezlini et al., 2022; Radin et al., 2020), and malaria (Chaudhury et al., 2022)).

Other fields have seen an increasing concentration of biomedical (Cook and Collins, 2015) and ML (Koch et al., 2021) research around pre-existing datasets (as opposed to generating and using novel datasets). In particular, some biomedical research has centered around pre-existing datasets from large-scale observational studies like *All of Us*[1] and the UK Biobank (Glynn and Greenland, 2020). These large-scale observational studies provide a diverse source of real-world

[1] https://www.researchallofus.org/publications/

human data that would be challenging for any research group to gather independently. Similarly, ML research is often organized around certain "benchmark" datasets. These benchmark datasets provide useful abstractions of certain tasks and serve as stable points of comparison between algorithmic implementations (Koch et al., 2021).

Given the increasingly central role pre-existing datasets play in biomedical and ML research, numerous studies have recently examined the generalizability of research findings across different datasets. Madigan et al. (2013) documented findings from clinical studies using cross-sectional observational datasets that do not generalize to other similar datasets. Similarly, ML models used for health applications (health-ML) often struggle to generalize to new datasets (Li et al., 2020; Johnson et al., 2018; Chekroud et al., 2024; Singh et al., 2022). Low generalizability is also well-known in more established ML disciplines (e.g., computer vision (Torralba and Efros, 2011), natural language processing (McCoy et al., 2019), and time series (Xu et al., 2022a)).

In light of persistent generalizability challenges, some studies have worked towards characterizing aspects of pre-existing datasets that lead to non-generalizable research. Research using "biased" datasets, or datasets with "unintended or potentially harmful" data properties, might be less generalizable (Vaughn et al., 2020). Dataset bias might stem from a combination of any number of distinct biases in data-generating processes. Furthermore, datasets gathered in observational studies (e.g., *All of Us* and the UK Biobank), are at a higher risk for systematic biases, like selection and information bias (Hammer et al., 2009). Some biases, such as representation bias along demographic axes, can lead to biased research (Wacholder et al., 2000; Abbasi-Sureshjani et al., 2020) but might be easier to mitigate. Other biases are likely harder to detect and account for. Any biases that impact the distributions of data underlying an ML model's training data might lead to poor generalizability in datasets without similar biases. Datasets are described as exhibiting "distribution shifts" if their underlying data are substantially different compared with another's (Cai et al., 2023).

Research using pre-existing datasets needs to be generalizable if it informs inferences

about the real world or develops ML models that might be deployed. However, generalizable research is particularly critical in the biomedical and health-ML domains, where outcomes might influence resource allocation or an individual's health outcomes. Our work was motivated by the observation that wearable data from pre-existing observational studies increasingly serve a dual research role[2]: as datasets for cross-sectional biomedical research and as benchmark datasets for developing health-ML models. However, the generalizability of findings from pre-existing, large-scale wearable device-based studies has not been previously examined. Our work aims to bridge this gap by (1) examining wearable data from multiple pre-existing, large-scale longitudinal wearable studies, (2) directly testing the generalizability of ML models on some existing monitoring tasks, and (3) examining the amount of performance change attributable to distribution shift (specifically, concept shift) in these datasets for these tasks.

## 4.2 Related Work

### 4.2.1 Demographic biases and associations in wearable datasets.

A large body of work examines demographic biases in large-scale wearable datasets or the association between certain demographics and wearable data. Schoeler et al. (2023) examined demographic biases in UK Biobank data and Doherty et al. (2017) found associations between wearable-measured accelerometry and demographics. Cho et al. (2022) demonstrated imbalances in *All of Us* FitBit data based on self-reported ethnicity along with several other bring-your-own-wearable device studies. Two studies have also used multiple regression in large, non-publicly available photoplethysmography-based heart rate (HR) datasets and both found that age, male sex, and white ethnicity were negatively correlated with mean HR (Golbus et al., 2021; Avram et al., 2019). Work on comparatively small ($<$100 participants), domain-focused, wearable-measured accelerometry datasets demonstrated bias in human activity recognition (HAR) datasets (Nair et al., 2023) and fall detection datasets (Casilari and Silva, 2022). However,

[2]https://allofus.nih.gov/news-events/announcements/research-roundup-all-us-participants-fitbit-data-drive-new-research

there has yet to be a comparison of the data underlying multiple longitudinal wearable datasets in conjunction with examining the impact of the previously documented demographic imbalances in these datasets.

### 4.2.2 Generalizability of wearable-ML models.

Broadly speaking, generalizable ML models perform similarly on data external to or different from their training data (Roelofs, 2019). Despite advancements, issues with model generalizability remain nearly ubiquitous across applied ML fields. The concept of generalizability remains largely unexplored in the wearable field, yet, limited research within the mobile health community has shown that ML models exhibit poor generalizability. Specifically, Adler et al. (2022) and Pillai et al. (2023) revealed that ML models using mobile phone data for passive mental health monitoring fail to generalize across studies. Xu et al. (2022b) similarly demonstrated that existing models trained to detect a specific chronic condition (depression) using mobile sensing data show poor generalizability across data gathered from the same study and site but in different years. To the best of our knowledge, no studies have examined the generalizability of longitudinal monitoring models across multiple wearable-based studies.

### 4.2.3 Distribution shifts.

The inability of ML models to generalize to external settings is commonly attributed to differences in the underlying distributions of data, called distribution shifts (Cai et al., 2023). Here, we assume distribution shift to be an umbrella term (as in Cai et al. (2023)) encompassing a few distinct types of shift. Consider data $(X, Y)$ with covariates (e.g., features) $X$ and labels $Y$ and a supervised learning model $f$ trained to predict $Y$ from $X$. $f$ might be applied to an external setting with data $(\tilde{X}, \tilde{Y})$ where distribution shifts can be decomposed into: label shift $p(Y)$ vs $p(\tilde{Y})$, covariate/feature shift $p(X)$ vs $p(\tilde{X})$, or concept shift $p(Y|X)$ vs $p(\tilde{Y}|\tilde{X})$. Many approaches that estimate label, covariate, or concept shifts assume at least one is held constant. However, in real-world data, all three types of shifts likely occur simultaneously. Indeed, acute

illness monitoring models deployed for surveillance (e.g., as in Radin et al. (2020)) are arguably deployed as *label shift* detection models. *Concept shift*, on the other hand, involves significant differences in the probability of certain outcomes within specific feature space boundaries across examples. For instance, if 0.5% of Americans with increased HR above a certain level had a viral infection, but 4% of Germans with the same increase in HR were infected, this might indicate a concept shift. Recent methodologies (Cai et al., 2023; Liu et al., 2023) were developed to quantify concept shift between datasets; we use their approach in these analyses. The most similar work with wearable data was performed by Vorburger and Bernstein (2006) using an entropy-based approach on short-scale accelerometry data. As far as we know, no studies have attempted to quantify concept shift between longitudinal wearable datasets.

## 4.3 Data

We sought datasets that were: (1) gathered using a commercially available wearable device capable of measuring HR (e.g., Apple Watch, FitBit, Oura Ring, etc.), (2) longitudinal (several weeks of data per participant on average), (3) large-scale (on the order of thousands of participants), and (4) labeled with timestamps that had not been anonymized in the time domain (e.g., shifted into future years, e.g., "2100"). Five datasets met these criteria: Homekit2020, GCD, COVID-RED, *All of Us*, and CDS. All but CDS had individually resolved wearable data with participant IDs (PIDs) linking their data to demographic information. All datasets had resting HR at daily-resolution except *All of Us*. We calculated *All of Us* daily resting HR directly using minute-resolution HR and step count data. Homekit2020, GCD, and COVID-RED all had daily questionnaires linked via PIDs which we used as ground truth labels for assessing acute illness monitoring model generalizability.

Table 4.1 summarizes the wearable device participants wore while in the study, the features available in each dataset, and the questionnaire outcomes used as ground truth labels for acute illness monitoring tasks. include further details on features, preprocessing steps, and

details on how to access each dataset.

**Table 4.1.** Descriptions of the datasets used in these analyses. See for participant counts expanded by demographics.

| Dataset | Device | Number of Participants | Demographics | Features | Questionnaires |
|---|---|---|---|---|---|
| HK | FitBit | n=5,012 | Sex, ethnicity, age, postal code | HR, activity, sleep | Symptoms, flu test results |
| TemPredict | Oura Ring | n=43,604 | Sex, ethnicity, age, education, etc. | HR, HRV, RR, sleep, activity, tempera-ture | Symptoms, COVID and flu test results |
| COVID-RED | Ava smartwatch | n=14,955 | Sex, ethnicity, age, education, BMI, etc. | HR, HRV, RR, temper-ature, perfusion index, sleep | Symptoms, COVID test results |
| *All of Us* | FitBit | n=13,735 | Sex, ethnicity, age, education, etc. | HR, activity, sleep | N/A |
| CDS | Any measuring HR | n=493,487 | N/A | HR, steps, sleep | N/A |

HK: Homekit2020, HR: heart rate, HRV: heart rate variability, RR: respiratory rate

### 4.3.1 Homekit2020

Homekit2020 was the first publicly available, large-scale wearable dataset wherein data from participants included demographic information, wearable data, and daily questionnaire data (Merrill et al., 2023). It includes FitBit data spanning December 2019 to April 2020 from over 5,000 adult participants recruited from across 50 U.S. states. Homekit2020 was also the first publicly available acute illness monitoring benchmark, and Merrill et al. (2023) trained and tested nine ML models on a set of acute illness monitoring tasks. For our results to be comparable, we attempted to reproduce their task definitions and training/testing procedures when examining model generalizability across datasets. See for more details.

### 4.3.2 Global COVID Dataset

The GCD includes Oura Ring data from January 2020 and through November 2020 from participants who owned an Oura Ring prior to the study and healthcare workers who were given an Oura Ring to participate in the study. Participants were distributed globally. Wearable device data, demographics, and daily questionnaires are available from over 40,000 participants. See for additional details.

### 4.3.3 COVID-RED

The COVID-RED dataset (Brakenhoff et al., 2023) includes Ava smartwatch data from February 2021 through November 2021 from over 14,000 adults living in the Netherlands along with demographics and daily questionnaires. Whereas the Homekit2020 and GCD datasets include minute-resolution wearable data, participants were instructed to wear the Ava bracelet only while asleep. Thus, COVID-RED wearable data is only provided at daily resolution and does not provide any notion of activity levels. These factors reduced the number of features shared between each dataset and without minute-level resolution data it was not feasible to test certain neural models as outlined in the Homekit2020 study. See Section 4.D for additional details.

### 4.3.4 *All of Us*

The *All of Us* research program is an ongoing major initiative to collect diverse health-related data, including electronic health records, genomic data, physical measurements, participant questionnaires, and wearable device data from over a million Americans (The All of Us Research Program Investigators, 2019). The *All of Us* research program emphasizes including groups typically underrepresented in biomedical research. The *All of Us* research program began allowing participants to share historical and prospective FitBit data starting in 2019. We use the *All of Us* Registered Tier Dataset v7. FitBit data is not paired with daily questionnaires at this time, thus we use these data for comparing resting HR distributions and not model generaliz-

ability. See Section 4.E for additional details, particularly how we calculated resting HR from minute-level data.

### 4.3.5 Corona-Dataspende

The CDS dataset (Wiedermann et al., 2023) includes geographically aggregated nightly mean values from over 400,000 adults from Germany. Data is available from April 2020 to December 2022. Data from any "fitness bracelet or smartwatch" from "Apple, Samsung, Fitbit, Garmin, Amazfit, Oura, Polar and Withings" were included in the dataset, and resting HR, steps, and sleep duration are available (Wiedermann et al., 2023). We used data aggregated across the entire nation of Germany to compare distributions of resting HR data with other datasets. See Section 4.F for further details.

## 4.4 Methods

We used these questions to guide our subsequent analyses:

1. What demographic biases exist in large-scale, longitudinal wearable datasets?
2. Are there substantial differences in the underlying data distributions even after statistically accounting for demographics?
3. How well can we expect acute illness detection models to generalize across wearable datasets when using community standard features and models?
4. How much of the changes in model performance across datasets is attributable to concept shift?

### 4.4.1 Demographic biases

Because there were substantial differences in the total number of participants in each dataset, we compared the proportion of participants in each demographic group to the proportions

in the U.S. population[3] and world population[4] (for age and sex), and the U.S. population for ethnicity. See Table 4.J.1 for the total numbers in each category.

### 4.4.2 Summarizing participant resting HR

Prior large-scale observational wearable studies aggregated all available wearable device data from each participant. As an example, Master et al. (2022) aggregated daily FitBit-measured step counts from each participant in the *All of Us* study and found that participants' average daily step count was correlated with incident disease (e.g., depression, hypertension, diabetes, etc.). For these analyses, we follow the approach taken in previous studies examining the relationship between demographic factors and real-world assessed HR (Avram et al., 2019; Golbus et al., 2021). Avram et al. (2019) performed a multiple linear regression and Golbus et al. (2021) performed an ANOVA (a special case of multiple regression (Nelson et al., 1979)) between several demographic factors and within-participant mean HR measurements. Our statistical approach was identical; however, fewer demographics were shared between these datasets (age, sex, and ethnicity) than those used in Avram et al. (2019) and Golbus et al. (2021). Our primary results focus on the mean daily resting HR as it is commonly used to assess acute illness.

### 4.4.3 Acute illness monitoring

We sought to use community standard methodological implementations to examine the performance and generalizability of acute illness monitoring models across datasets. Therefore, we reviewed nineteen prior acute illness monitoring studies to determine community standards (see Section 4.A for criteria and Tables 4.L.1 to 4.L.3 for results). Thirteen trained ML models on longitudinal wearable data for acute illness monitoring. In the cases where there was no obvious community standard, we attempted to reproduce methodological approaches taken in the Homekit2020 study wherever feasible.

[3]https://www.census.gov/data/tables/2020/demo/age-and-sex/2020-age-sex-composition.html
[4]https://genderdata.worldbank.org/topics/population/

**Ground truth definitions**

Prior acute illness monitoring studies have a wide range of ground truth definitions (see Table 4.L.2 for a summary). Given the absence of obvious community standards surrounding ground truth labels, we follow the approach taken by the Homekit2020 of "one prediction per participant per day" as suggested by Nestor et al. (2023). Any day without missing wearable data in the nights leading up to a ground truth label from a daily questionnaire was used for evaluating the performance of our models (see Section 4.N and our code for details). We work with three of the tasks described in the original Homekit2020 study: prediction of respiratory viral infection (confirmed by laboratory test), flu symptoms[5], and fever symptoms (see Sections 4.B to 4.D for details on how labels were extracted from each dataset). In other words, we set no minimum wearable device or questionnaire compliance levels to include a participant's data in these analyses except that we required enough data within a rolling baseline period (at least six of ten days) to reliably calculate the mean and standard deviation of their wearable data.

**Normalization strategy**

Eleven of the thirteen acute illness monitoring studies we reviewed used a lagged, within-individual z-score normalization (Table 4.L.1). The other studies also used a lagged baseline approach; Quer et al. (2022) used the median and inter-quartile range as opposed to mean and standard deviation, and Risch et al. (2022) performed an unspecified lagged baseline normalization. There seems to be community consensus around the use of within-individual, lagged baseline normalization, however, no two studies chose the same combination of baseline window length (the number of days used to calculate the mean and standard deviation in the baseline period) and window offset (the number of days the normalization period is from the ground truth day). Therefore, we performed a hyperparameter grid search on window length and offset to determine an optimal normalization strategy based on these datasets. Implementation details are shown in Section 4.M and we found that z-scoring by a ten-day window with a

[5]We used the more common Centers for Disease Control and Prevention definition of flu symptoms

twelve-day offset was optimal for these data.

**Feature set**

Prior reviews have examined the features used by these models and their performance (Mitratza et al., 2022). To test generalizability, we considered the set of features shared between the Homekit2020, GCD, and COVID-RED datasets: 1) resting HR and 2) time spent asleep. In order for our results to be comparable to the Homekit2020 study, we focused on prediction tasks as they did. Thus, the input features into our model included the three days of z-score normalized resting HR and time spent asleep (as described in Section 4.4.3) prior to a ground truth day (see Section 4.M for details). We also one-hot encoded the day of the week so that models could account for human activities that follow seven-day rhythmicity (e.g., work days). We note that models tend to perform better on detection tasks (using data from up to the night after a ground truth day, Section 4.N) and that within-dataset performance is lower when limiting features to those that are shared across datasets (HR and sleep vs all available features, Section 4.N).

**Model choice**

Boosting, tree-based classifiers (e.g., XGBoost, LGBM, Sklearn's gradient boosting classifiers) are commonly used in many acute illness monitoring studies (Merrill et al., 2023). In our review of acute illness monitoring studies (Table 4.L.3), we found that a plurality of studies reported results from at least one boosting, tree-based classifier. Because we aimed to use common community implementations, we chose to use Sklearn's histogram gradient boosting classifier (Pedregosa et al., 2011). See Section 4.M for implementation details.

**Evaluation metrics**

We examined model performance using the area under the receiver operating curve (AUROC), which is commonly reported in acute illness monitoring studies. Despite its bias in situations with extreme class imbalance, AUROC allowed us to calculate meaningful percentage changes when comparing models trained on one dataset and tested on the other datasets. We also

considered using average precision (AP), however, the relative changes as assessed by AP did not result in meaningful percentages of change. Performance as described by AP are shown in Table 4.N.3.

**Training and testing**

The Homekit2020 study found that models performed about as well on a "user split" (respective to "time split") when following a train-test cross-validation procedure; we use a modified version of their approach for within-dataset performance evaluation. When evaluating under the user split setting, a model is trained on data from one group of participants and tested on another. Given the extreme class imbalance in these data, we implemented a stratified version of Homekit2020's user split to ensure that each train-test split had a similar number of participants with positive examples. For within-dataset performance, reported metrics represent the average across a five-fold randomly stratified user cross-validation split. To assess generalizability, models were trained on all available data from one dataset and tested on all available data from each of the other datasets.

### 4.4.4 Performance change due to concept shift

We used recently developed methods (*WhyShift*) to estimate the proportion of performance change due to concept shift (Cai et al., 2023; Liu et al., 2023). Their method takes a trained model from one dataset, test data from the same dataset, and an external dataset as input. It uses a domain classifier to estimate a subset of examples in the test data and external dataset that have features with shared support. It then uses these examples with shared support to estimate the performance change that can be attributed to concept shift. See Section 4.O for implementation details and a schematic further describing how *WhyShift* estimates performance changes due to concept shift.

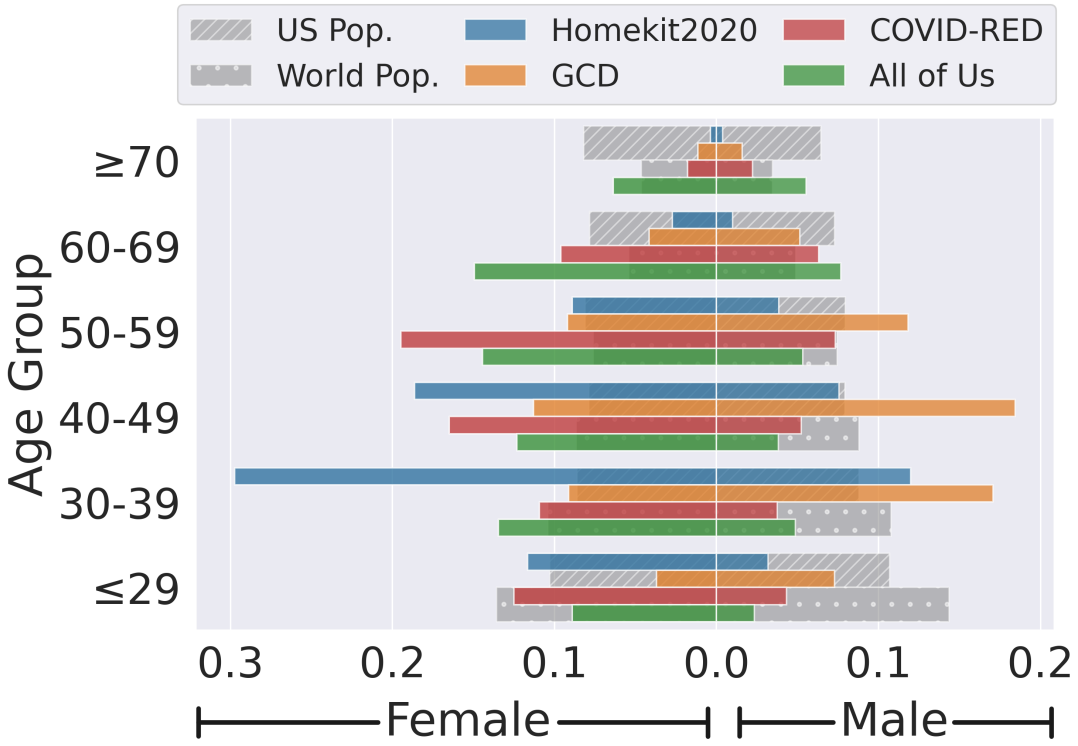


**Figure 4.1.** Datasets are biased in self-reported age and sex, relative to both the U.S. and World populations. Within-dataset participant counts are normalized by the total number of participants in each dataset and displayed using a population pyramid.

## 4.5 Results

These analyses suggest that large-scale wearable datasets are substantially biased based on the relative prevalence of self-reported age, sex, and ethnicity. We found opposite directional correlations between age and resting HR and significant differences in mean resting HR in each dataset. Most models performed worse on external datasets. The majority of performance changes could be attributed to concept shift.

### 4.5.1 Demographic biases

Each dataset is substantially biased based on the relative prevalence of self-reported demographics. These large-scale wearable studies tend to be over-representative of younger and female groups (Figure 4.1) as well as White groups (Figure 4.2).

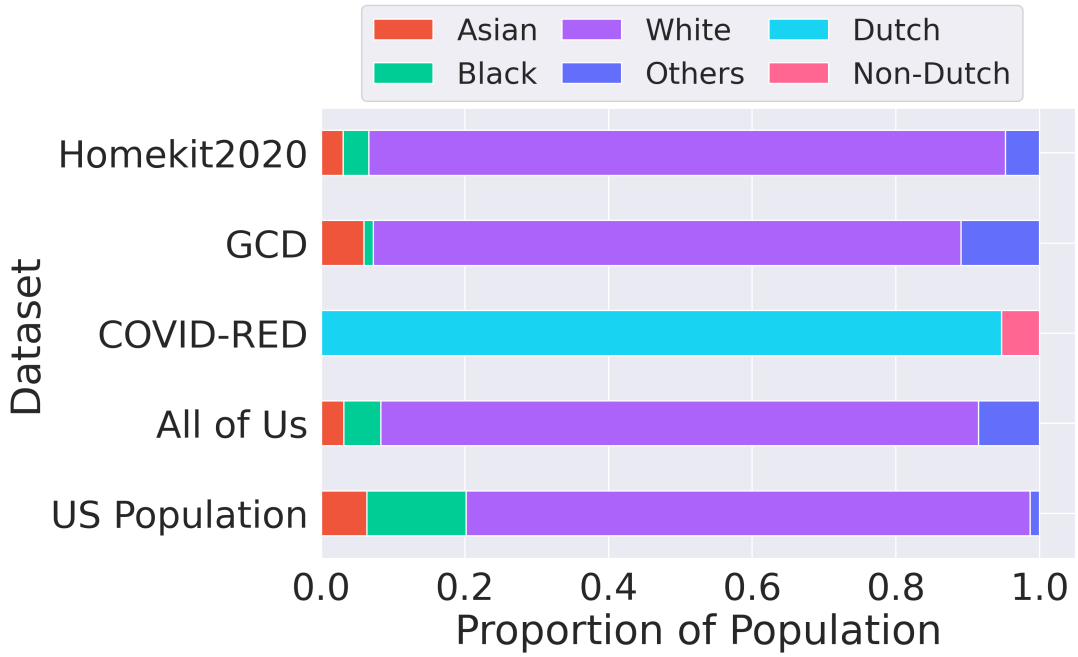


**Figure 4.2.** Datasets are biased in self-reported ethnicity as compared to the U.S. population, particularly with respect to Black participants. Within-dataset participant counts are normalized by the total number of participants in each dataset and are displayed based on the relative prevalence of self-reported ethnicity.

### 4.5.2 Average dataset resting HR

There appear to be substantial differences in the underlying distributions of within-dataset average resting HR (Figure 4.3) and a variety of within-dataset trajectories throughout the year which might correspond to changes in behavior in the U.S. during the COVID-19 lockdown in. We also provide visualizations of the minute-of-day means for HR and activity split by age, sex, and ethnicity for the Homekit2020, GCD, and *All of Us* datasets in Figures 4.F.1 to 4.F.3 along with descriptions of weekday vs. weekend differences across datasets (Section 4.H).

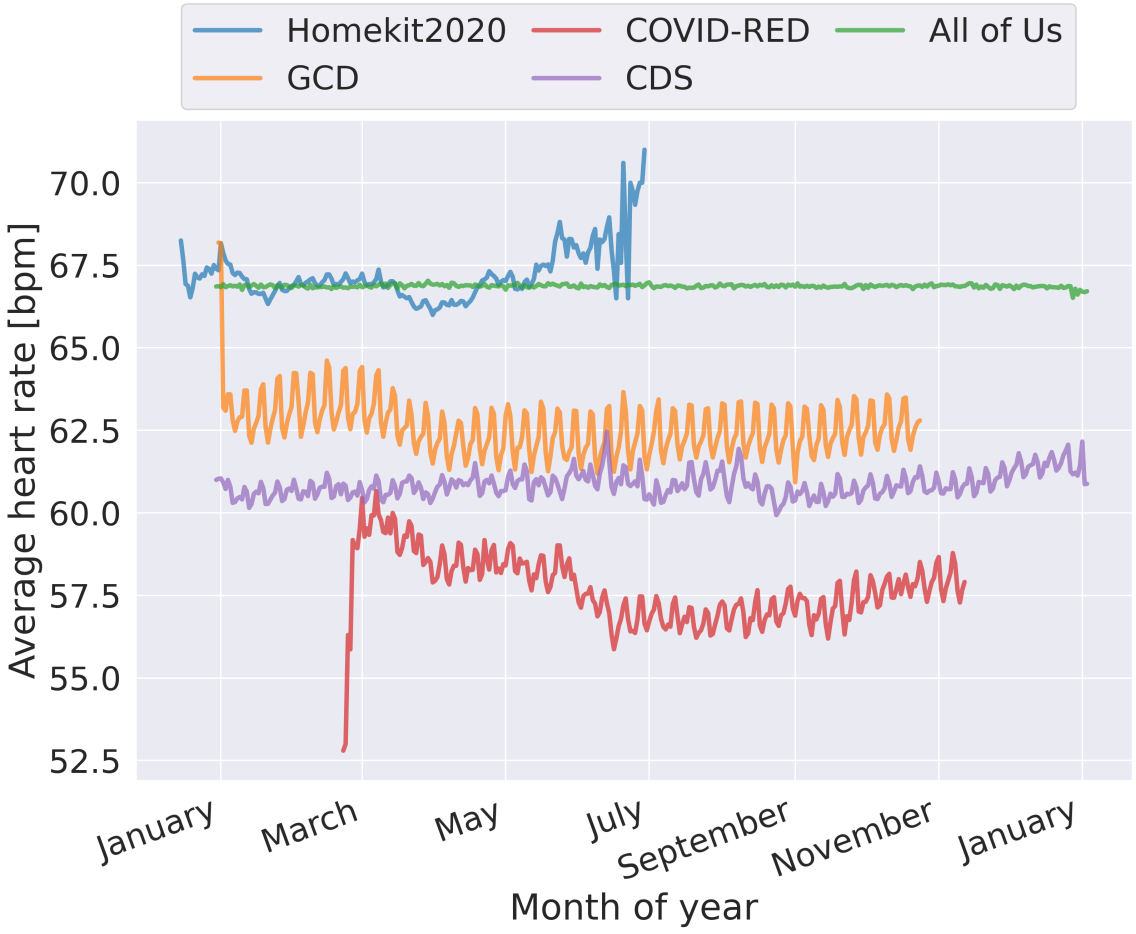


**Figure 4.3.** Within dataset mean resting HR varies substantially between datasets. Here, the average daily resting HR was taken as the mean across all participants with available data on the same relative date (i.e., 2nd Tuesday of each year) and the mean across repeated relative dates for datasets spanning multiple years (*All of Us* and CDS).

### 4.5.3 Within-dataset HR differences

Regardless of dataset, when accounting for age, sex, and ethnicity, males tend to exhibit lower HRs than females (Table 4.2) and African-American participants exhibit higher HRs relative to white participants. Notably, age is positively correlated with HR in the GCD dataset, while age is negatively correlated with HR in the *All of Us* dataset.

**Table 4.2.** Results are from a multiple regression with age, sex, and ethnicity as factors/covariates and mean resting HR as response values. Values are reported as: regression coefficient (p-value). Datasets exhibit concordant correlations for mean resting HR vs sex and a subset of ethnicities, however, the correlation between age and HR is conflicting between datasets.

| **Dataset** | **Age** | **Sex*** | **Ethnicity†** | | |
|---|---|---|---|---|---|
| | | | Black | Asian | Other |
| Homekit2020 | -0.014 (0.226) | -3.56 (<0.001) | 2.363 (<0.001) | -0.095 (0.886) | 0.937 (0.082) |
| GCD | 0.059 (<0.001) | -3.21 (<0.001) | 4.125 (<0.001) | 0.657 (<0.001) | 0.062 (0.585) |
| COVID-RED | 0.082 (0.045)** | -3.52 (<0.001) | Non-Dutch: 0.881 (0.001) | | |
| *All of Us* | -0.108 (<0.001) | -4.69 (<0.001) | 5.495 (<0.001) | -0.685 (0.125) | 1.502 (<0.001) |

*Reference: female, †Reference: Caucasian/white, **Coded as discrete bins of 10 years vs continuous

**Table 4.3.** Within-dataset performance (bold) is the mean AUROC across five-fold cross-validation. Models tested on external data are trained on all internal data. "Mean others": mean within-task AUROC on external data. "Percent drop": change between within-dataset performance and "Mean others."

| *Task* | Train \ Test | Homekit2020 | GCD | COVID-RED | Mean others | Percent drop |
|---|---|---|---|---|---|---|
| *Viral* | Homekit2020 | **0.780** | 0.496 | 0.586 | 0.54 | 30.64 |
| | GCD | 0.534 | **0.565** | 0.510 | 0.52 | 7.61 |
| | COVID-RED | 0.705 | 0.508 | **0.588** | 0.61 | -3.15 |
| *Flu* | Homekit2020 | **0.620** | 0.641 | 0.654 | 0.65 | -4.44 |
| | GCD | 0.613 | **0.673** | 0.689 | 0.65 | 3.27 |
| | COVID-RED | 0.568 | 0.620 | **0.685** | 0.59 | 13.28 |
| *Fever* | Homekit2020 | **0.701** | 0.628 | 0.666 | 0.65 | 7.7 |
| | GCD | 0.679 | **0.673** | 0.694 | 0.69 | -2.01 |
| | COVID-RED | 0.653 | 0.630 | **0.685** | 0.64 | 6.35 |

**Table 4.4.** The majority of performance changes are attributable to concept shift. Values represent the proportion (concept:total) of performance change attributable to concept shift. Results displayed are the mean across a five-fold cross-validation, with the test dataset from cross-validation used with external data to estimate the performance changes due to shifts.

| *Task* | Train \ Test | Homekit2020 | GCD | COVID-RED |
|---|---|---|---|---|
| *Viral* | Homekit2020 | - | 0.91 | 0.94 |
| | GCD | 0.98 | - | 1.0 |
| | COVID-RED | 0.99 | 1.0 | - |
| *Flu* | Homekit2020 | - | 0.77 | 0.55 |
| | GCD | 0.76 | - | 0.72 |
| | COVID-RED | 0.613 | 0.74 | - |
| *Fever* | Homekit2020 | - | 1.0 | 0.77 |
| | GCD | 1.0 | - | 0.73 |
| | COVID-RED | 0.81 | 0.78 | - |

### 4.5.4 HR differences across datasets

Multiple regression confirms the qualitative assessment observed in Figure 4.3: with respect to the *All of Us* dataset, participants from the Homekit2020 dataset have the most similar HRs (1.90 bpm lower on average), followed by participants from the GCD dataset (4.86 bpm lower) and the COVID-RED dataset (10.41 bpm lower, Table 4.K.1). When pooling all participants across all datasets, males still tend to have lower HRs than females (3.69 bpm lower) while age was positively (but not significantly) correlated with HR (0.001 bpm/decade).

### 4.5.5 Acute illness monitoring generalizability

In general, performance was worse on external datasets, however, this was not always the case (Table 4.3). Across all tasks, the average performance drop was 6.58%. Prediction of viral positivity seemed substantially easier for examples in the Homekit2020 dataset respective to the GCD and COVID-RED datasets. Models trained on the GCD dataset (the largest dataset by number of training examples) exhibited the lowest average drop in performance (2.96%) across all tasks on external datasets respective to Homekit2020 (11.3%) and COVID-RED (5.49%). Indeed, when testing on COVID-RED data for the flu and fever tasks, models trained

on GCD data marginally outperformed models trained on COVID-RED data. Models trained on the Homekit2020 and GCD datasets both performed better on the flu symptom task in the COVID-RED dataset relative to their within-dataset performance.

### 4.5.6 Concept shift drives performance differences

These analyses suggest that the overwhelming majority of performance changes between datasets were due to concept shift (Table 4.4). The proportion of performance change attributable to concept shift was also approximately symmetric for each task (e.g., trained on COVID-RED, tested on GCD was close to trained on GCD, tested on COVID-RED). The viral positivity task exhibited the highest average concept shift proportion at 0.97. Flu exhibited the lowest concept shift proportion at 0.69.

## 4.6 Limitations

This study has several limitations. First, unknown or unmeasured confounding variables might explain the observed differences in correlations between age and resting HR. Such differences might stem from unaccounted-for dataset biases or the non-ergodic nature of these measures (Mangalam et al., 2023). Additionally, these data were gathered in different years and some data might reflect changes due to the onset of COVID-19 in early 2020 rather than typical human physiology. Furthermore, we considered the within-participant mean of resting HR across time, which likely compressed much of the time-dependent information in these data (e.g., menstrual cycles). Future work could explore these time-dependent characteristics (e.g., with autoregressive models) and examine differences between datasets. These datasets were gathered using different wearable devices and prior work suggests that FitBit devices might *underestimate* HR respective to gold-standard reference HR measurements (Fuller et al., 2020). These results, however, suggest that participants in the FitBit-utilizing Homekit2020 and *All of Us* datasets had *higher* average HRs. We stress that the intention of these analyses is not to claim that any of these datasets or devices used therein provide a more accurate

representation of reality, but rather that researchers examining any one of these datasets in isolation could come to wildly different results and thus interpretations of reality. Prior research also documents such dataset-dependent discrepancies (Madigan et al., 2013). Our normalization strategy reduces differences in within-dataset means (Section 4.P). Future work might examine whether such within-individual normalization strategies mitigate dataset biases, however, our specific normalization strategy might only be optimal for these datasets and tasks.

We evaluated the performance of a single, albeit effective and widely utilized, classifier and we did not explore whether domain adaptation techniques enhance model generalizability or mitigate concept shifts or whether features other than resting HR vary between datasets. We do not intend to surmise that the models we used are at the forefront of acute illness monitoring technology. Nonetheless, their performance is comparable to the best-performing models in the original Homekit2020 study. Differences in model performance might be attributable to our use of an optimized, community standard baseline normalization technique. Future work could also consider other architectures, particularly deep neural networks, which we were unable to examine due to substantial differences in the sampling structure of each dataset. We did not consider the transferability of model hyperparameters across datasets and tasks; this would be an important future step in developing more generalizable models and might prove even more important in work with deeper architectures. Similarly, researchers could explore whether domain adaption approaches (e.g., Fernando et al., 2013; Singh, 2021) improve model generalizability. *Unsupervised* domain adaptation approaches might be particularly promising for these datasets - given the large number of unlabelled examples - and in deployment where labels might not be immediately available. Unsupervised domain adaptation approaches also might perform well under concept shift scenarios (Rostami and Galstyan, 2023), which remains largely unexplored in wearable datasets. These analyses offer a baseline against which to compare the impact of implementing such methods. Similar comparative analyses have not been performed for large-scale accelerometry data (e.g., those available in the *All of Us* and UK Biobank studies, though work currently under review uses multiple such datasets (Shim et al., 2023)). We found

this surprising given that the body of literature correlating accelerometry measures from these datasets with health conditions or using these data to develop ML models is much larger than the illness monitoring literature. Future work similar to ours could consider the accelerometry data underpinning the *All of Us* and UK Biobank studies.

## 4.7 Conclusion

Given the time and expense required to collect large-scale wearable datasets, it would not be surprising if researchers performing cross-sectional observational studies or developing ML models for health monitoring tasks coalesced around a few of the pre-existing wearable datasets. At least, similar dataset concentration occurred in many of the more established ML communities. The *All of Us* and UK Biobank datasets are emerging as the default large-scale wearable datasets. Nestor et al. (2023) caution attention to the study design and outcomes described in acute illness monitoring studies, and the original authors of the Homekit2020 study (Merrill et al., 2023) suggest that performance on any of these datasets is not indicative of real-world performance. Our work underscores that such caution is merited, and we suggest that wherever possible, future studies involving these datasets should test whether correlations and models generalize across other large-scale datasets.

Ultimately, the data from large-scale wearable device-based studies show impressive utility in describing human physiology, especially as it changes over time, and might be useful to develop and train ML models for monitoring tasks. Indeed, in cases where the results from these data are used for resource allocation, like in epidemiological settings, even small improvements can save lives. Such applications are particularly promising given that millions of people already own and use wearable devices that are connected via their mobile device to the internet, potentially enabling improved resource allocation in near real-time. However, to the extent that a community of researchers forms around these large-scale datasets and works towards developing models for acute illness detection, we hope that this work serves as a reminder that these datasets

likely face many of the same challenges known all too well by other research communities. In the case that acute illness monitoring using wearables continues to develop into a more established health-ML field, we hope this work spurs a discussion around anticipating and accounting for the biases and generalizability challenges documented here.

# Appendix

## 4.A Studies reviewed for community standards

In order to determine community standard acute illness monitoring approaches, we reviewed the same studies (Bogu and Snyder, 2021; Cleary et al., 2022; Hassantabar et al., 2020; Hirten et al., 2021; Lonini et al., 2021; Miller et al., 2020; Mishra et al., 2020; Natarajan et al., 2020; Nestor et al., 2021; Quer et al., 2022; Shapiro et al., 2021; Smarr et al., 2020) as those in a previous review of the performance of wearable devices for the detection of SARS-CoV-2 (COVID-19) (Mitratza et al., 2022). We also manually supplemented these studies with studies that were published after this review was published and we also included studies focused on acute illnesses other than COVID-19 (e.g., flu). We prioritized reviewing other studies that focused on acute viral respiratory diseases and found: (Alavi et al., 2022; Mayer et al., 2022; Conroy et al., 2022; Risch et al., 2022; Quer et al., 2021; Dunn et al., 2022). We excluded studies that did not use a commercially available wearable device (e.g., those only available as a medical device (Goldstein et al., 2021) or based on custom hardware (Zhang et al., 2021; Kumar et al., 2023)) or if it was not clear what device was used (Lakshmi and Robinson Joel, 2023). We also excluded studies that were limited to small, non-representative sub-populations (e.g. children ages 3-17 who had recently received an appendectomy (Ghomrawi et al., 2023), patients undergoing chemotherapy for gastrointestinal cancer (Low et al., 2017)) or non-human research subjects (Davis et al., 2021). Furthermore, we did not consider protocol publications (Larimer et al., 2021) or publications that were not peer-reviewed (Skibińska, 2023). We also found several studies that focused on illnesses that were either not acute or not respiratory (e.g.,

chronic inflammatory rheumatic disease (Rao et al., 2023), stress (Miyawaki et al., 2023), or Parkinson's disease (Li et al., 2023)).

## 4.B Homekit2020 Dataset

Homekit2020 is a dataset provided by researchers at the University of Washington and Evidation and this study recruited adult participants from across 50 U.S. states and includes data from December 2019 to April 2020. It was the first publicly available, large-scale wearable dataset wherein data from participants included demographic information, wearable data (FitBit; activity, heart rate, and sleep), and responses to daily questionnaires. In their original publication, Merill et al. provide a set of acute illness monitoring tasks and implement and test nine ML models, which they use to demonstrate state-of-the-art performance on these tasks. Here we describe the details of the task definitions, data processing steps, and training/testing procedures that we use to test the generalizability of acute illness monitoring models across datasets.

- Data access: Data from this study is available to "qualified researchers" who agree to the study's "Conditions for Use". Researchers need to have a user profile through the Synapse platform and are required to submit an "Intended Data Use" statement in order to access these data. Data is available from Synapse.

- Code access: Code defining Homekit2020's original models, data preprocessing, and data loaders are available at this GitHub repository. The code used for the analyses in this work is available at this GitHub repository.

- Features: Prior to the start of the study, participants owned a FitBit device capable of measuring steps, sleep, and heart rate. Inclusion criteria included residency in the U.S., the ability to read, speak, and understand English, no diagnosis of flu in the 3 months before the start of the study, willingness to complete a daily online questionnaire for the study's duration, ownership of an iPhone, iPad, or Android smartphone or tablet, readiness to

download an app if experiencing flu-like symptoms, willingness to complete an at-home flu test kit and send the sample to a laboratory using a pre-paid shipping label. Daily averages, including resting heart rate, were calculated by FitBit and retrieved using the FitBit API. Features include: resting heart rate, minutes spent in bed, sleep efficiency, the number of naps, the total time spent asleep, the total time in bed, the number of calories burned doing activities the previous day, the total number of calories burned the previous day, the number of calories burned by an individual's basal metabolic rate, the total marginal estimated calories burned for the day, the number of: sedentary, lightly active, fairly active, and very active minutes from the previous day.

- Labels: During the study period, participants were asked to complete a daily, online questionnaire. Responses to this questionnaire are provided in the "daily_surveys_onehot.csv" file available on Synapse. This questionnaire included questions about symptoms and self-reported temperature among other questions. The questionnaire for symptoms was based on severity using a four-point Likert scale. Results from a comprehensive initial questionnaire and PCR diagnostic tests were included as separate tables (initial questionnaires are found under the "2020_04_30" folder on Synapse and PCR results are in the "lab_results_with_triggerdate.csv") and participants are linked across tables via PIDs. If a participant indicated experiencing ILI symptoms in the daily questionnaire, they were then given additional follow-up questionnaires. These subsequent questionnaires were more detailed and aimed to gather more information about their symptoms. In cases where symptoms were reported, participants were directed to self-administer a flu test. This test would provide immediate results for a generic influenza infection. The test sample was also meant to be sent to a laboratory for a more detailed analysis to determine the specific type of virus, if any. We reviewed the original Homekit2020 publication (Merrill et al., 2023), another study from the same authors using the Homekit2020 dataset (Merrill and Althoff, 2022), an earlier publication from Evidation (Kolbeinsson et al., 2021), and the code from

Merrill et al. (2023) available at https://github.com/behavioral-data/Homekit2020 to determine how the group created ground truth labels. For symptom-based labels (flu and fever), ground truth labels were generated using participants' responses to daily questionnaires. For fever, if a participant reported experiencing a severe fever "defined as three or more on a four-point Likert scale" that day was labeled positive. In the original Homekit2020 study, the flu task was described as "Will the participant report two or more flu symptoms (including cough, fever, and fatigue) of any severity today?" On the other hand, the original flu monitoring study (Kolbeinsson et al., 2021) does not include fatigue in the list of symptoms and states that a day was labeled positive for flu symptoms if a participant reported: "two specific symptoms (cough and one of body ache, feeling feverish, chills, sweats) on the same day". Given the lack of consensus both in these studies and their published code, we opted to implement a more common definition of flu symptoms, which is also the definition used for influenza-like illness surveillance from the CDC: "fever or feverishness plus either cough or sore throat" [6]. We took the same approach for the GCD and COVIR-RED datasets. It was *not* explicitly stated in either the original publications or their code how the authors defined negatively labeled examples. However, we found that selecting days wherein participants completed the symptom questionnaire *and* did not experience these levels of symptoms produced class balances close to the results reported in (Merrill and Althoff, 2022). For viral positivity, we found that labeling all days except for those wherein a participant reported testing positive by a PCR test to produce class balances most similar to those reported in (Merrill and Althoff, 2022). We used these approaches for labeling examples in the GCD and COVID-RED studies.

- Demographics: Participant demographics are linked by participant IDs that can be found under the "PublicPortal\homekit2020_export_2020_04_30" folder in the "screener" files on Synapse.

[6] https://www.cdc.gov/quarantine/air/management/guidance-cruise-ships-influenza-updated.html

- Acknowledgments: These data were contributed by participants as part of the Home Testing of Respiratory Illness Study developed by Evidation Health and described in Synapse (doi.org/10.7303/syn22803188).

## 4.C Global COVID Dataset

The Global COVID dataset (GCD) was gathered as part of a larger study by researchers at several R1 research institutions in collaboration with Ōura Health Oy. Participants were recruited on a rolling basis from individuals who already owned an Oura Ring and at healthcare sites at over 20 different healthcare institutions throughout the U.S. Participants who already owned an Oura Ring were distributed globally. Participants were recruited starting in March of 2020 and recruitment stopped in September 2020. Data was back-filled for participants who already owned the device and data is available from January 2020 to November 2020. Wearable device data, demographics, and daily questionnaires are available from over 40,000 participants.

- Data access: We obtained access to the dataset through a data-use agreement that does not allow the data to be made publicly available.

- Code access: Code for processing the Global COVID dataset directly is not available, however, we used processing functions that were identical to those used for the Homekit2020 and COVID-RED datasets and examples from these datasets are available at this GitHub repository.

- Features: Summary values ("sleep summaries") from when a participant was asleep include: resting heart rate, the lowest heart rate from the sleep period, heart rate variability (rMSSD), respiratory rate, respiratory rate variability, temperature deviation from a user's long-term temperature average, temperature trend deviation from a three-day rolling average, sleep onset latency, time spent awake, time spent in REM sleep, time spent in light sleep, time spent in deep sleep, and time spent asleep. Sleep summaries were calculated by the device and retrieved by the researchers using Oura's API.

- Labels: During the study period, participants were asked to complete a daily, online questionnaire. This questionnaire included questions about symptoms including: fever,

sore throat, dry cough, cough with mucus, and cough with blood. We combined dry cough, cough with mucus, and cough with blood into a single ”cough” label. The questionnaire for symptoms was binary (experienced or did not experience). As outlined in Section 4.B, if participants reported fever and either cough or sore throat, that day was included as a positive example in the flu task. During their time in the study, participants were also asked to report if they tested positive for any respiratory viral illnesses (COVID-19, flu). We used responses to these questions for the viral positivity task.

- Demographics: Participants completed a baseline questionnaire wherein they reported certain demographic information including age, sex, and ethnicity. Baseline questionnaire data is linked to the participants' wearable and questionnaire data via PIDs.

## 4.D COVID-RED Dataset

The COVID-RED dataset was gathered as part of the COVID-RED study, a collaboration between nine organizations: UMC Utrecht, Ava, Julius Clinical, University College London, the Danish Center for Social Science Research, Sanquin, Takeda, Roche, and Dr Risch. Adults from the Netherlands were recruited starting in February 2021 and data is available through November 2021. Wearable device data (Ava smartwatch), demographics, and daily questionnaires are all available, however, whereas the Homekit2020 and GCD datasets include minute-resolution wearable data, participants were instructed to wear the Ava bracelet only while asleep. Thus, COVID-RED wearable data is only available at daily resolution and does not provide any notion of activity levels.

- Data access: Data is publicly available from Dataverse.
- Code access: To the best of our knowledge, the code used in the studies by the authors who gathered the COVID-RED data is not publicly available. The code used for the analyses in this work is available at this GitHub repository.
- Features: The study aimed to enroll a total of 20,000 subjects, focusing on residents of the Netherlands. To be eligible, participants needed to be at least 18 years old and residents of the Netherlands. They were required to own a smartphone compatible with the study requirements (running at least Android 8.0 or iOS 13.0) and be able to read, understand, and write Dutch. Individuals were excluded if they had a previous positive test for SARS-CoV-2 (either through PCR/antigen or antibody tests), were currently suspected of having a coronavirus infection or exhibiting symptoms, had an electronic implanted device (like a pacemaker), or suffered from cholinergic urticaria. Participants were recruited from previously studied cohorts and through public campaigns. Interested individuals were directed to visit the COVID-RED web portal. Here, they completed questionnaire questions to determine their eligibility and expressed their interest in joining the study.

After completing the questionnaire and indicating their interest, eligible participants received a subject information sheet and a consent form. Their enrollment was confirmed upon compliance with the study's inclusion and exclusion criteria and after providing consent. Enrolled subjects were instructed to complete the Daily Symptom Diary in the Ava COVID-RED app, wear the Ava bracelet each night, and synchronize it with the app daily for the duration of the study. Wearable measured features are available in the "wd_20230515.csv" file and are labeled by the date they were gathered. Since these data were gathered at night, we confirmed whether the labeled date corresponds to data from the night before or the night after the labeled date by taking the mean across all points from the same day of the week and looking for known weekly rhythms. This confirmed that these data were from the night before the date. Wearable measured features include: resting heart rate ("WDPULSE"), respiratory rate ("WDRESP"), skin temperature ("WDTEMP"), heart rate variability ("WDPULSEV"), perfusion index ("WDOXI"), and total time spent asleep ("WDSLEEP").

- Labels: Participants were asked to complete a Daily Symptom Diary. This was facilitated through the Ava COVID-RED app, a specially designed application for this study. The app was to be installed on the participants' smartphones, which had to be compatible with the app's requirements. Each day, participants were prompted to report their health status and any symptoms they might be experiencing. Within the "wd_20230515.csv", under the "WDSYMP" column, reported symptoms are comma-separated. We used responses in this column labeled as "no_current_symptoms" as our negative class label across tasks. Examples were included as positive class examples for the fever task if "fever" was included in the list of symptoms. If "fever" and either "cough", or "sore_throat" were reported we included that example in the flu task. For the viral positivity task, we used the "WDDIAG" column and labeled examples with "positive" as positive.

- Demographics: Participant demographics are linked by a participant ID and can be found

in “dm_20230515.csv”. Note, that we included the country of birth provided in this dataset as ethnicity (i.e., Dutch vs. non-Dutch) as this was the closest proxy to ethnicity that was available from the COVID-RED dataset. This might not be directly comparable to the concepts of race/ethnicity used in the U.S. and the Homekit2020, GCD, and *All of Us* datasets.

## 4.E *All of Us* Dataset

The *All of Us* research program is a major initiative to collect diverse health-related data, including electronic health records, genomic data, physical measurements, participant questionnaires, and wearable device data from over a million Americans. It emphasizes including groups typically underrepresented in biomedical research. The *All of Us* research program began allowing participants to share historical and prospective FitBit data starting in 2019. As of January 2024 (*All of Us* Registered Tier Dataset v7), FitBit data in *All of Us* are not paired with any daily questionnaires at this time.

- Data access: Data used in this study are from the *All of Us* Registered Tier Dataset v7. Researchers from institutions with a Data Use and Registration Agreement in place with *All of Us* can create an account. After identity confirmation, completion of the mandatory training, and signing the data user code of conduct, researchers can begin to access Registered Tier data. Data is then accessible through an online service that provides compute for a fee. Researchers at qualified institutions can register at https://www.researchallofus.org/register/

- Code access: The code used for the analyses in this work is available at this GitHub repository.

- Features: *All of Us* participants who already owned a FitBit could consent to share their wearable device data with the *All of Us* research program. Minute resolution steps and heart rate are available along with activity summaries. Because this dataset does not provide a FitBit-calculated resting heart rate (as was provided in (Merrill et al., 2023)), we calculate one using the approach outlined in Alavi et al. (2022), taking the mean of any available minute-resolution heart rate values between the hours of midnight and 7 AM local time when, in the same minute (matched by day, hour, minute, and participant ID), the number of FitBit measured steps is 0. See the SQL query defined in our code for

how this was calculated, which is available in "all_of_us_analyses.ipynb" at this GitHub repository.

- Labels: N/A

- Demographics: Demographic information is linked in the *All of Us* database via participant IDs. Age was not explicitly provided so it was calculated using participants' provided date of birth referenced to January 1st, 2019, which is when participants began sharing FitBit data. See "all_of_us_analyses.ipynb" available at this GitHub repository for further details on querying the *All of Us* database for these demographics.

- Acknowledgments: The *All of Us* Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the *All of Us* Research Program would not be possible without the partnership of its participants.

## 4.F Corona-Dataspende Dataset

The Corona-Dataspende dataset resulted from a collaboration between the Robert Koch Institute and Humboldt University of Berlin. Adults from Germany were recruited to donate their wearable device data starting in April 2020; data collection ended in December 2022. Data from any "fitness bracelet or smartwatch" from "Apple, Samsung, Fitbit, Garmin, Amazfit, Oura, Polar and Withings" were included in the dataset and the publicly available version of the dataset is aggregated across geographic regions. The dataset is available as the mean across all participants with available data for a particular night. These means are calculated across varying levels of geographical aggregation. We used data aggregated across the entire nation of Germany to compare distributions of resting heart rate data from other datasets.

- Data access: This dataset is publicly available and can be downloaded directly from Zenodo.
- Code access: The code used for the analyses in this work is available at this GitHub repository.
- Features: Participants included anyone over 16 with access to a German app store. Over a million participants downloaded the app, with more than 500,000 individual participants contributed at least one data point from a wearable. Regular participation in questionnaire studies involved up to 30,000 people. Data includes mean daily resting heart rate, step count, and sleep duration, aggregated by geographical units based on European NUTS (NUTS3 to NUTS0) classifications. Data is available from April 2020 to December 2022. Data are spatial averages, which prevents identifying any single individual's data. Data is excluded from users with incomplete postal codes, Apple Watch sleep data, and implausible vital signs. Any data point with more than 50,000 steps per day, more than 24 hours of sleep, or with a resting heart rate below 30 or above 150 beats per minute was excluded.

- Labels: N/A
- Demographics: Individual-level demographic information is not available.
- Acknowledgments: N/A

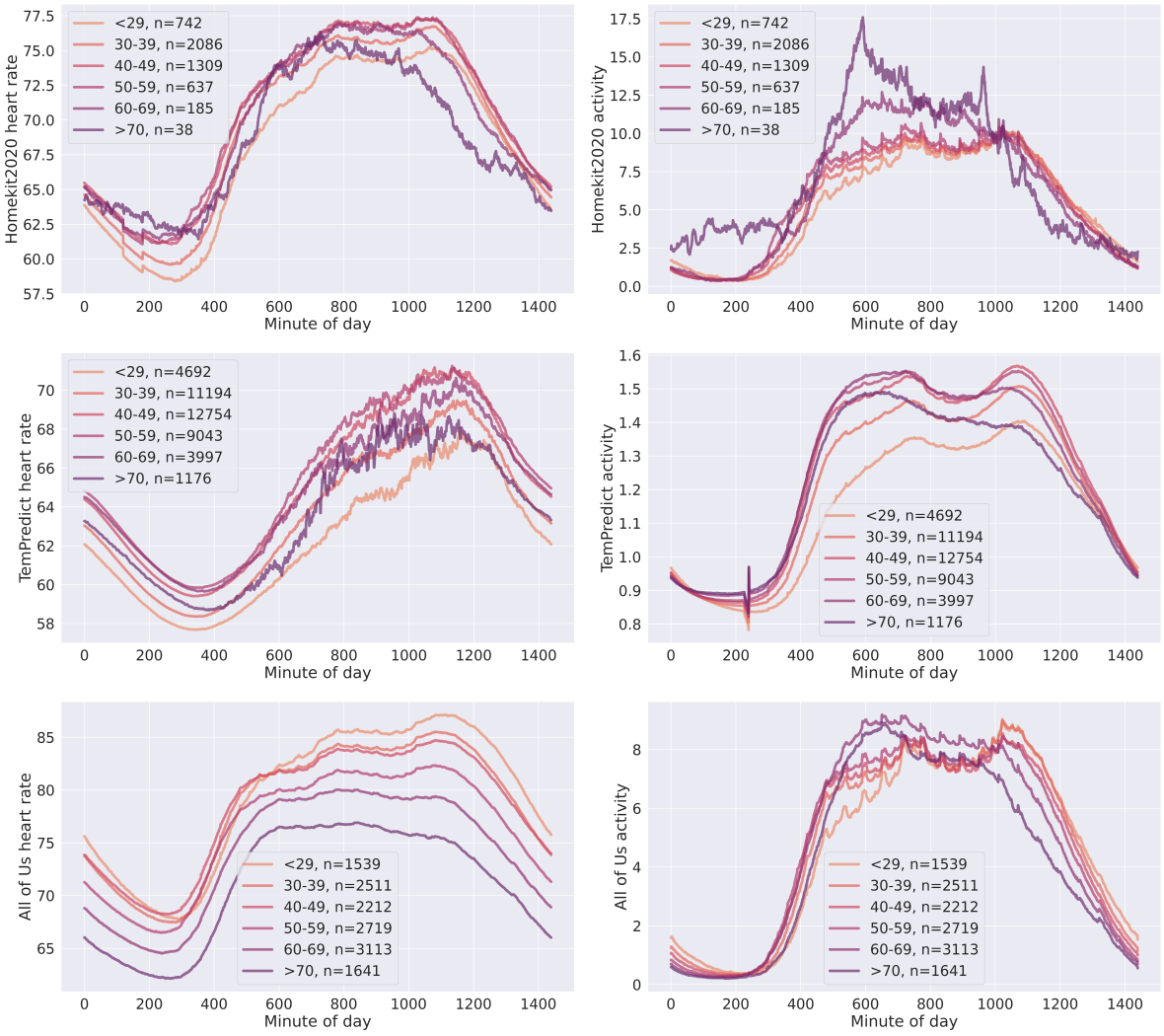


**Figure 4.F.1.** There are within-dataset differences in the mean resting HR based on age throughout the entire day. There also appear to be differences in the patterns of HR and activity throughout the day when comparing across datasets. Lines represent the mean time-of-day wearable-measured average HR (left) and wearable-measured activity (right). Here, we stratify participants by age and take the within-dataset mean (top: Homekit, middle: GCD, bottom: *All of Us*) for each age group using all available data from that minute of the day.

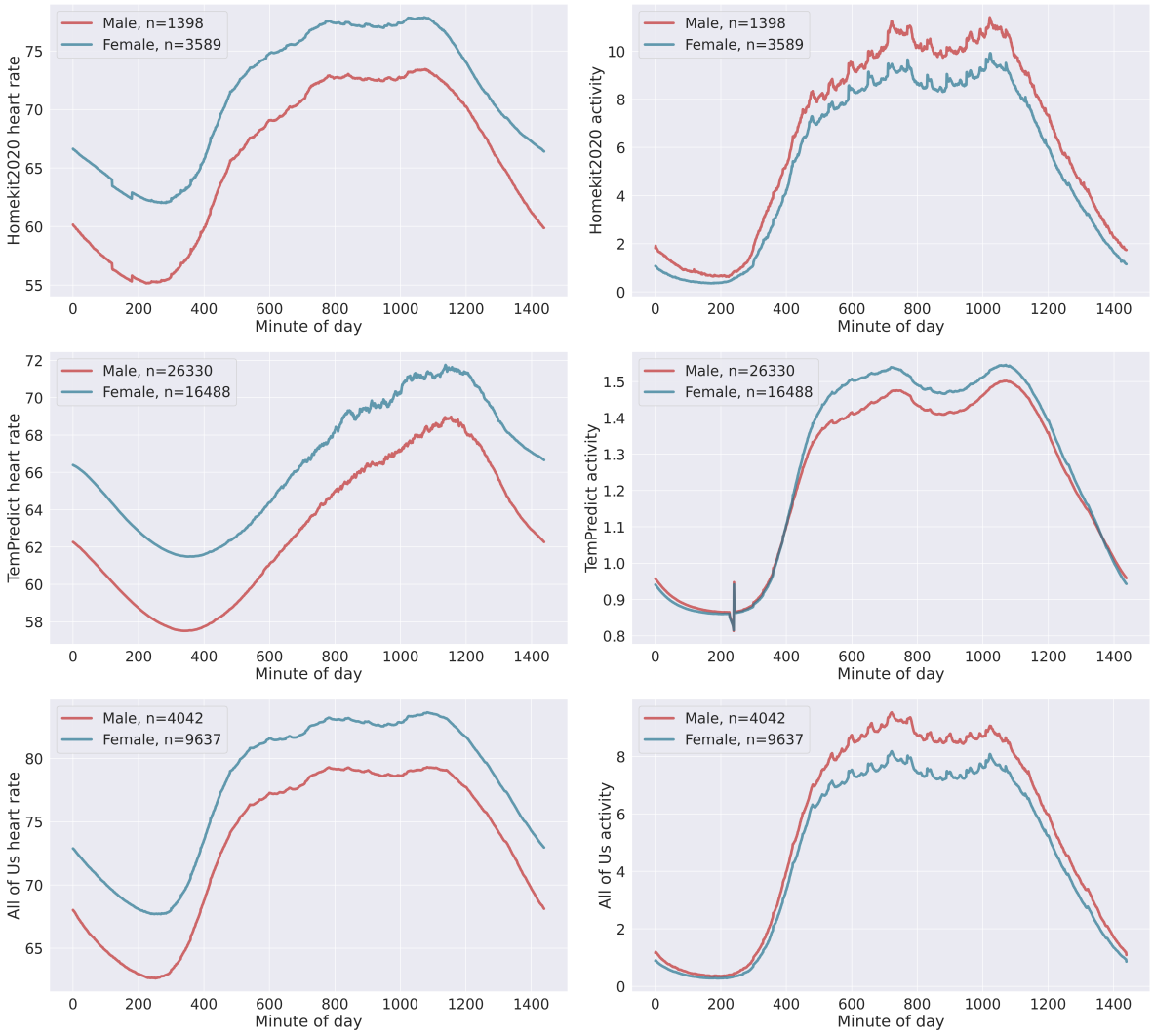


**Figure 4.F.2.** There are within-dataset differences in the mean resting HR based on biological sex throughout the entire day. There also appear to be differences in the patterns of HR and activity throughout the day when comparing across datasets. Time-of-day wearable-measured average heart rate (left) and wearable-measured activity (right). Here, we stratify participants by sex and take the within-dataset mean (top: Homekit, middle: GCD, bottom: *All of Us*) for each sex using all available data from that minute of the day.

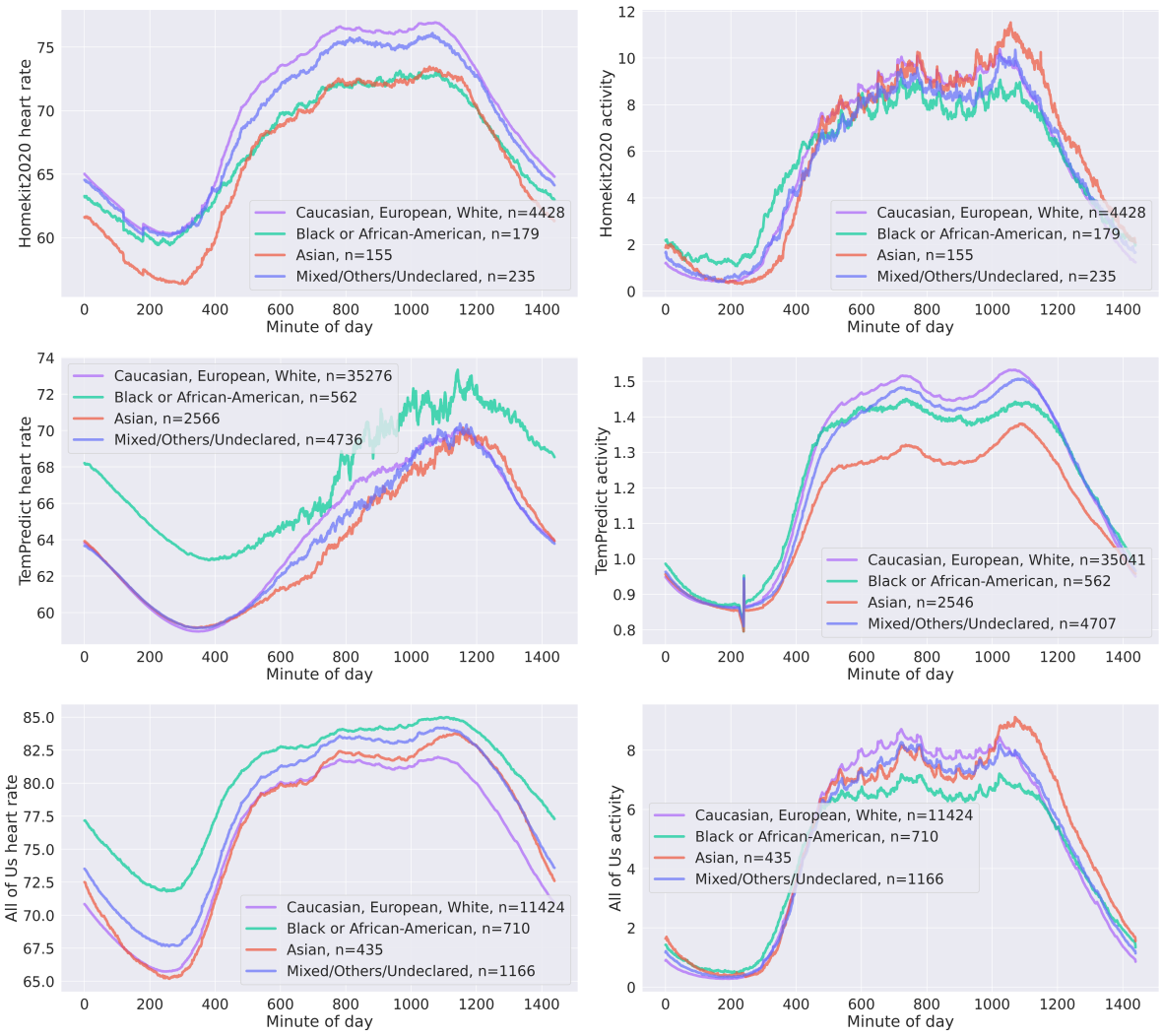


**Figure 4.F.3.** There are within-dataset differences in the mean resting HR based on ethnicity throughout the entire day. There also appear to be differences in the patterns of HR and activity throughout the day when comparing across datasets. Time-of-day wearable measured average heart rate (left) and wearable measured activity (right). Here, we stratify participants by ethnicity and take the within-dataset mean (top: Homekit, middle: GCD, bottom: *All of Us*) for each ethnicity using all available data from that minute of the day.

## 4.G Dimensionality reduction of participant level data

Bias can manifest in a number of ways. To properly measure a dataset's bias, we would need to compare wearable device measured heart rate with a clinical "gold standard", measured in the same participant concurrently. However, none of these datasets have such data. A proxy exists for a subset of the All of Us participants, for which we have access to FitBit data and clinical heart rate measurements from those participants. Note, however that these lab measurements are biased in their own way given that they were measured during a visit to a clinic as opposed to the "free-living" physiological measurements from their FitBit devices. Regardless, mean resting FitBit measured heart rate is significantly correlated with mean EHR reported heart rate (Pearson correlation coefficient=0.504, $p<0.001$). The mean absolute error between mean resting FitBit measures heart rate and mean EHR reported heart rate is 7.13 (STD=10.18) bpm.

As demonstrated Table 4.2, resting heart varies based in age, sex, and ethnicity. In order to further contextualize the extent to which such variability manifests as observable differences between datasets, we used unsupevised learning to describe summary statistics of all a participant's available data. Since the total number of days of data varies by individual, we limited our summary statistics to those that are length invariant and tolerant to missingness. Accordingly, we calculated the first 4 statistical moments of each time series (mean, standard deviation, skew, and kurtosis), along with the median, variance, the variation coefficient to quantify some notions of the distributions of resting heart rate values. Furthermore, to quantify some notion of time-dependent information contained in these data, we calculated the first 20 autocorrelation coefficients and parameters from an SARIMA model fit to each participant's data. The autocorrelation coefficients capture some of the structured variance on the day scale across time and perform particularly well at separating participants with strong weekly seasonality. An SARIMA model is an ARIMA (auto regressive, integrated, moving average model) with an additional seasonal component (Nobre et al., 2001). These are models are quite common in time series analysis and forecasting (Nobre et al., 2001). We determined the optimal SARIMA model

parameters by performing a hyperparameter grid search on the Germany dataset with AIC as our search criteria. We used the Germany dataset because 1) it does not have participant-level data so individual-level comparisons would not be possible and 2) we assume this dataset to be our best approximation of a true distribution because it is the daily mean from by far the greatest sample size available, (well over a hundred thousand participants). We found that the lowest error model was achieved using an SARIMA with p=3, d=1, q=3, P=2, D=1, Q=1, and s=2. We then embed each of the features using UMAP into low (i.e., two) dimensional space to qualitatively assess the similarity of individuals between datasets. This embedding is shown in Figure 4.G.1 as a scatter plot. In general, individuals from the same dataset tend to cluster together. Furthermore, based on the distance between the center of mass for each dataset (the average embedding across participants from the same dataset), it seems that the Homekit dataset and the COVID-RED dataset are more similar to each other than the All of Us dataset.

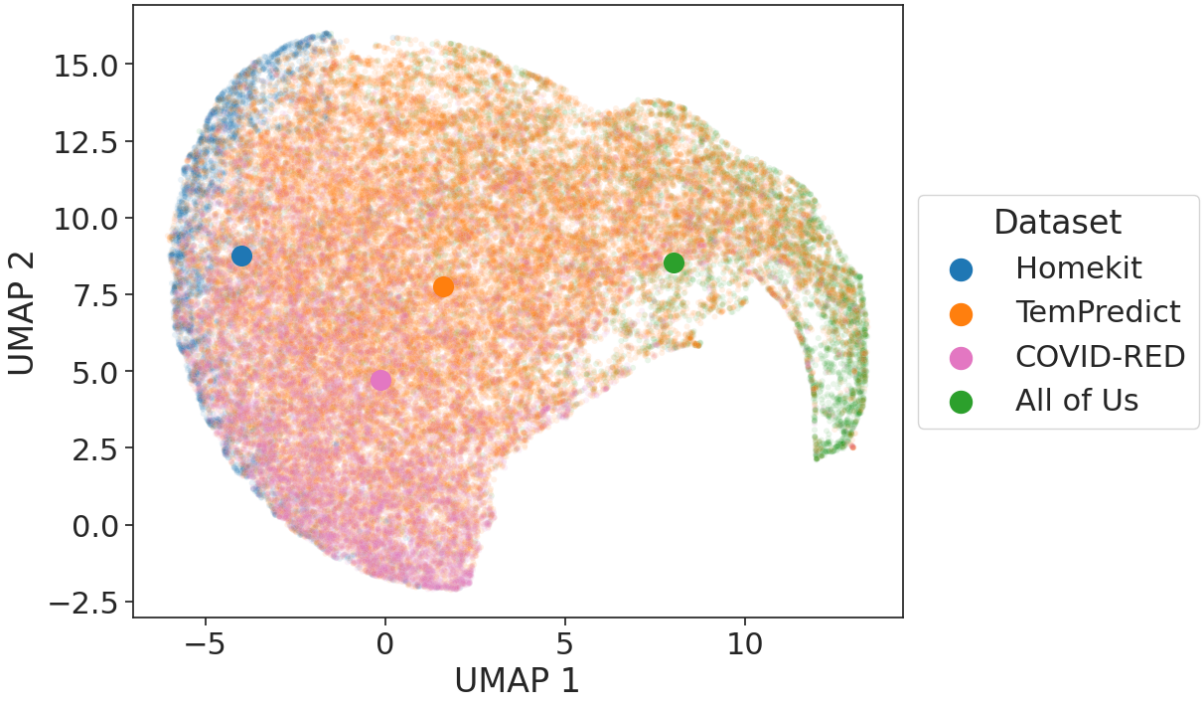


**Figure 4.G.1.** Low-dimensional embeddings of features derived from participants' longitudinal resting heart rate. Each point in the UMAP scatter plot is from a single participant and is colored by the dataset that participant was from.

## 4.H Weekday vs weekends differences by dataset

We observed differences in the mean values observed on weekend nights (Friday night or Saturday night) compared to weeknights (all other nights). The difference in means (weekend effect, WE) between these two sets of nights (weeknight vs weekend) varied by dataset. The largest WEs were observed in the GCD (1.20 beats per minute, bpm) and COVID-RED (0.62 bpm) datasets. WEs were less pronounced in the CDS dataset (0.39 bpm) and seemingly absent in the *All of Us* dataset (0.01 bpm).

## 4.I Data preprocessing

Conservative reasonableness bounds were used to filter the resting heart rate (HR) and time spent asleep features. Resting HR measurements below 20 bpm or above 200 bpm were set to NaNs and excluded from subsequent analyses. Time spent asleep measurements below 60 seconds or above 16 hours were similarly set to NaNs and excluded from subsequent analyses.

## 4.J Participant counts

**Table 4.J.1.** The total number of participants from each dataset whose data are used in Figures 4.1 to 4.3 and Tables 4.2 and 4.K.1

| | **Homekit2020** | | **GCD** | | **COVID-RED** | | ***All of Us*** | |
|---|---|---|---|---|---|---|---|---|
| Age bin* | Male | Female | Male | Female | Male | Female | Male | Female |
| <29 | 160 | 584 | 3115 | 1607 | 646 | 1870 | 319 | 1218 |
| 30-39 | 601 | 1491 | 7368 | 3984 | 561 | 1635 | 665 | 1842 |
| 40-49 | 379 | 934 | 8032 | 4970 | 784 | 2468 | 522 | 1685 |
| 50-59 | 193 | 446 | 5184 | 4061 | 1096 | 2915 | 728 | 1975 |
| 60-69 | 50 | 136 | 2241 | 1844 | 942 | 1438 | 1051 | 2045 |
| >70 | 19 | 19 | 691 | 507 | 334 | 266 | 757 | 872 |
| Ethnicity | | | | | | | | |
| Asian | 155 | | 2587 | | Dutch: 14167 | | 435 | |
| Black or African-American | 179 | | 571 | | | | 710 | |
| Caucasian, European, White | 4450 | | 35718 | | Non-Dutch: 788 | | 11424 | |
| Mixed/Others/Undeclared | 238 | | 4766 | | | | 1166 | |

*Does not include sex reported as "other". Train/test data for models were not filtered by demographics

# 4.K Inter-dataset differences

**Table 4.K.1.** Men have significantly lower resting heart rates in a pooled samples across datasets and there are significant differences between datasets in mean resting heart rate. Results are from a multiple regression was with age bin, sex, and dataset as factors/covariates and mean heart rate as response values. Age bins are based on decades as in Table 4.J.1. Reported as: regression coefficient (p-value).

| **Age bin** | **Sex*** | **Dataset†** | | |
|---|---|---|---|---|
| | | Homekit | GCD | COVID-RED |
| 0.001 (0.961) | -3.69 (<0.001) | -1.90 (<0.001) | -4.86 (<0.001) | -10.41 (<0.001) |

*Female as reference, †*All of Us* as reference

## 4.L Feasibility study review

**Table 4.L.1.** Here, we examine the normalization techniques and exclusion criteria used by nineteen studies. While all the acute illness monitoring feasibility studies that use machine learning (ML) approaches use a lagged baseline normalization, there does not appear to be a community consensus around the window size and offset used for this normalization.

| Study | Baseline start | Baseline end | Min. # days | Normalization |
|---|---|---|---|---|
| Bogu and Snyder 2021 | Unclear | Unclear | N/A | Z-score |
| Cleary et al. 2022 | -21 | -7 | 7 | Median/IQR |
| Hassantabar et al. 2020 | Not longitudinal | Not longitudinal | Not longitudinal | Min-max scaling |
| Hirten et al. 2021 | Not ML | Not ML | Not ML | Z-score |
| Lonini et al. 2021 | Not longitudinal | Not longitudinal | N/A | N/A |
| Miller et al. 2020 | -30 | -14 | N/A | Z-score |
| Mishra et al. 2020 | -28 | -1 | N/A | Z-score |
| Natarajan et al. 2020 | -5 | 0 | N/A | Z-score |
| Nestor et al. 2021 | -35 | -7 | 14 | Z-score |
| Quer et al. 2022 | -21 | -7 | N/A | Median/IQR |
| Shapiro et al. 2021 | Not ML | Not ML | Not ML | Not ML |
| Smarr et al. 2020 | Not ML | Not ML | Not ML | Not ML |
| Alavi et al. 2022 | -7 or -28 | -1 | N/A or 14 | Z-score |
| Mayer et al. 2022 | -35 | -8 | 1 | Z-score |
| Conroy et al. 2022 | -17 | -7 | 5 | Z-score |
| Risch et al. 2022 | -28 | -10 | 29 consecutive | "baseline normalization" |
| Merrill et al. 2023 | -7 | -1 | 5 | Z-score |
| Quer et al. 2021 | -21 | -7 | N/A | Z-score |
| Dunn et al. 2022 | -60 | -22 | 19 | Z-score |

**Table 4.L.2.** Here we examine how nineteen studies chose to define their positive ground truth and negative ground truth examples. Acute illness monitoring feasibility studies seemingly have wildly different task definitions.

| Study | Positive ground truth | Negative ground truth |
|---|---|---|
| Bogu and Snyder 2021 | -7 to +21 relative to symptom onset | -10 to -20 relative to symptom onset |
| Cleary et al. 2022 | 0 to +7 days after symptom onset | -21 to -7 days prior to symptom onset |
| Hassantabar et al. 2020 | Not longitudinal | Not longitudinal |
| Hirten et al. 2021 | N/A | N/A |
| Lonini et al. 2021 | Not longitudinal | Not longitudinal |
| Miller et al. 2020 | Days -2 days prior to symptom onset to +3 | -30 to -14 days prior to symptom onset |
| Mishra et al. 2020 | -14 to +7 days relative to symptom onset | N/A |
| Natarajan et al. 2020 | +1 to +7 days after symptom onset | -21 to -8 days prior to symptom on set |
| Nestor et al. 2021 | Symptom start to symptom end | All other |
| Quer et al. 2022 | -21 to -7 relative to symptom onset | 0 to +7 relative to symptom onset |
| Shapiro et al. 2021 | Not ML | Not ML |
| Smarr et al. 2020 | Not ML | Not ML |
| Alavi et al. 2022 | 21 days before the symptom onset for symptomatic cases or diagnosis date for asymptomatic cases or -28 | 21 days before a negative test result, the entire time frame for untested participants, or days before the detection window for positive participants |
| Mayer et al. 2022 | -7 to 14 days around COVID symptom onset | -35 to -8 days before COVID symptom onset |
| Conroy et al. 2022 | -14 to -1 days prior to a positive COVID test | -14 to -1 days prior to a negative COVID test |
| Risch et al. 2022 | -2 days prior to symptom onset | -20 to -3 days prior to symptom onset |
| Merrill et al. 2023 | -1 days prior to symptom onset | Not explicitly stated |
| Quer et al. 2021 | +1 to +7 after symptom onset | -21 to -7 days prior to symptom onset |
| Dunn et al. 2022 | -5 to -1 days prior to symptom onset | -60 to -22 days prior to symptom onset |

**Table 4.L.3.** Here, we examine which models were used by nineteen studies. Acute illness monitoring feasibility studies employ a wide variety of models and architectures, however, a plurality chose to use a variation of gradient boosting tree-based classifier.

| Study | Model used in study |
|---|---|
| Bogu and Snyder 2021 | LSTM-based autoencoder |
| Cleary et al. 2022 | Not ML |
| Hassantabar et al. 2020 | Deep neural network |
| Hirten et al. 2021 | Not ML |
| Lonini et al. 2021 | Logistic regression |
| Miller et al. 2020 | Gradient boosted classifier |
| Mishra et al. 2020 | Finite state model, Isolation Forest |
| Natarajan et al. 2020 | Neural network |
| Nestor et al. 2021 | XGBoost and Gated recurrent units |
| Quer et al. 2022 | Logistic regression |
| Shapiro et al. 2021 | Not ML |
| Smarr et al. 2020 | Not ML |
| Alavi et al. 2022 | Finite state model, Isolation Forest |
| Mayer et al. 2022 | Linear SVM |
| Conroy et al. 2022 | Gradient boosting ensemble learning method |
| Risch et al. 2022 | LSTM |
| Merrill et al. 2023 | XGBoost, CNN, Transformers, ResNet |
| Quer et al. 2021 | Multivariate logistic regression |
| Dunn et al. 2022 | Logistic regression, K-nearest neighbor, support vector machine, random forest, and extreme gradient boosting |

## 4.M Hyperparameter tuning and model configuration

In order to determine the optimal window offset and window length, we treat each as hyperparameters to be tuned and thus performed a grid search over window offset and window length. We used a logistic regression (LR) model trained on a pooled sample of a random, equal number of participants from each dataset and found the average AUROC across each task (prediction of testing positive for a respiratory virus, flu symptoms, and fever symptoms). Other hyperparameters were left at default Sklearn settings. LR models used resting heart rate and time spent asleep from the night before the ground truth day as features. We found that the best performance occurred when z-scoring by a ten-day window with a twelve-day window offset, where the offset is the number of days the baseline period is away from the normalized day.

We used Sklearn's (v1.2.0) Histogram-Based Gradient Boosting Classification Tree (sklearn.ensemble.HistGradientBoostingClassifier) as our primary model. We pooled a sample of examples together across each dataset and task and performed a hyperparameter search over a range of hyperparameters and found that models performed and generalized well with early stopping disabled, a learning rate of 0.1, l2 regularization at 0.2, and the rest of the hyperparameters left at default. For model comparisons, we chose to use the three days leading up to a ground truth day as it balanced model overfitting against having fewer examples to train and test with.

## 4.N Prediction vs detection

On average, models performed better on the detection version of each task (a model operating on data from nights before and after a ground truth day) as compared with the prediction version of each task (a model operating on data from the nights strictly before a ground truth day) on the same dataset. We tested this by training models on the same ground truth labels using the normalization strategy described in Figure 4.M.1 (ten-day window length with a twelve-day window offset). The prediction model included z-score normalized data from Nights -3, -2, and

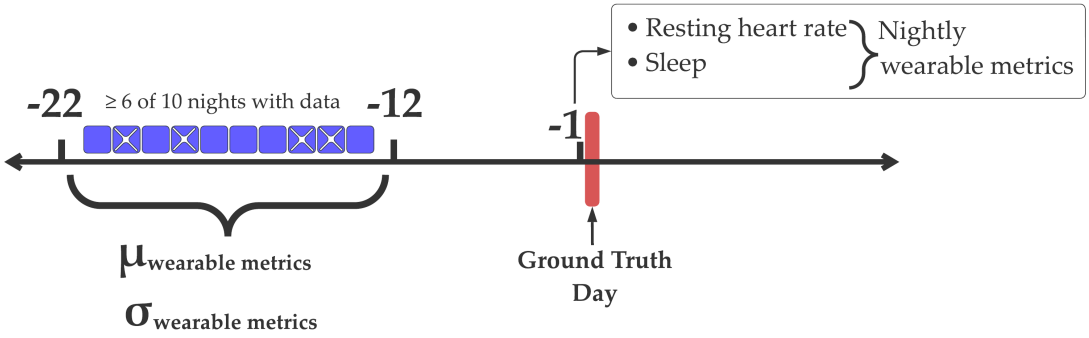


**Figure 4.M.1.** Schematic of the optimized baseline z-score strategy showing an example of how wearable data from the night before the ground truth day is normalized. For detection, data from the night after the ground truth day is z-score normalized by a window that is also shifted forward by one night so that its window is still lagged by twelve days.

-1 relative to the ground truth day. The detection model included z-score normalized data from Nights -2, -1, and 0 (0 being the first night after a ground truth day). Thus, the total number of features was held constant within datasets (one value for each feature for each night), however, the timing of those features was changed. Models for testing prediction and detection are based on all available nightly features, features used for these models are described in the dataset descriptions in Sections 4.B to 4.D. Training and testing follow a stratified five-fold user split cross-validation schema as described in Section 4.4.3.

**Table 4.N.1.** Performance on prediction tasks across datasets.

| *Task* | Metric \ Dataset | Homekit2020 | GCD | COVID-RED |
|---|---|---|---|---|
| *Viral* | AUROC | 0.858 | 0.592 | 0.628 |
| | AP | 0.0020 | 0.0017 | 0.0014 |
| *Flu* | AUROC | 0.637 | 0.713 | 0.657 |
| | AP | 0.0159 | 0.0287 | 0.0306 |
| *Fever* | AUROC | 0.766 | 0.742 | 0.686 |
| | AP | 0.0363 | 0.0857 | 0.0955 |

**Table 4.N.2.** Performance on detection tasks across datasets.

| *Task* | Metric \ Dataset | Homekit2020 | GCD | COVID-RED |
|---|---|---|---|---|
| *Viral* | AUROC | 0.931 | 0.592 | 0.638 |
| | AP | 0.0112 | 0.0017 | 0.0041 |
| *Flu* | AUROC | 0.638 | 0.713 | 0.700 |
| | AP | 0.0160 | 0.0287 | 0.0479 |
| *Fever* | AUROC | 0.770 | 0.734 | 0.709 |
| | AP | 0.0159 | 0.0845 | 0.0998 |

**Table 4.N.3.** Performance of the shared-features model on detection tasks across datasets as measured by average precision (AP).

| *Task* | Train \ Test | Homekit2020 | GCD | COVID-RED |
|---|---|---|---|---|
| *Viral* | Homekit2020 | 0.0007 | 0.002 | 0.0007 |
| | GCD | 0.00018 | 0.0013 | 0.00057 |
| | COVID-RED | 0.0001 | 0.002 | 0.0011 |
| *Flu* | Homekit2020 | 0.0118 | 0.0029 | 0.0173 |
| | GCD | 0.010 | 0.0064 | 0.019 |
| | COVID-RED | 0.007 | 0.0024 | 0.033 |
| *Fever* | Homekit2020 | 0.0066 | 0.0093 | 0.058 |
| | GCD | 0.0039 | 0.019 | 0.049 |
| | COVID-RED | 0.006 | 0.0086 | 0.068 |

# 4.O *WhyShift* implementation

Liu et al. (2023) implemented a method for estimating the proportion of performance change that can be attributed to concept shift $Y|X$ and covariate shift $X$ (note, we follow their notation here). Their results rely on the DISDE method, originally outlined in Cai et al. (2023). If data $(X,Y)$ from a training distribution $P$ are used to train a classifier $f$ and $f$ is to be used on some target distribution $Q$, then $P$ and $Q$ have some shared support $S$, which they estimate using an auxiliary domain classifier $\hat{\pi}$ trained to differentiate between examples in $P$ and examples in $Q$. The DISDE method then estimates the performance of $f$ trained on $P$ on examples in $S$ and $Q$. It uses the performance of $f$ on $P$, $S$, and $Q$ to estimate performance changes due to $Y|X$ shifts and $X$ shifts. In this case, $X$ shifts take the form of $P \rightarrow S$ shifts and $S \rightarrow Q$ shifts. We use *WhyShift's* implementation of the DISDE framework as it handles training the domain classifier and decomposes the performance changes due to distribution shifts. For implementation, we used the same stratified, user-split cross-validation for training models. We passed the model ($f$) trained on the training split in each cross-validation as the input model to *WhyShift*, the test split from cross-validation for examples from $P$, and all examples from each other dataset for examples from $Q$. We report the average performance change due to concept shift across cross-validation splits. We also added the same histogram gradient-boosting classifier for the domain classifier $\hat{\pi}$ as it was not originally implemented in *WhyShift*.

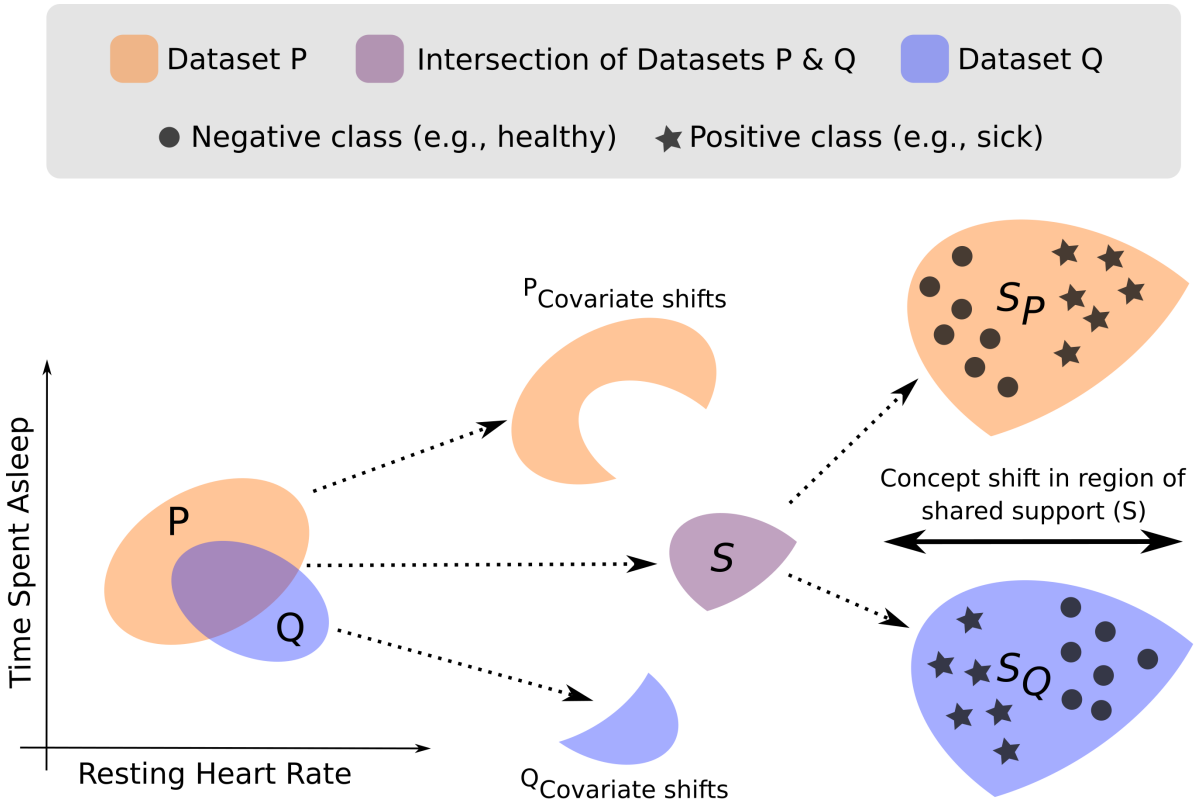


**Figure 4.O.1.** Datasets (distributions) $P$ and $Q$ might both exhibit concept shift and covariate shift. Concept shift can only be estimated in the subsections of feature space which have shared support (e.g., have overlap in their distributions) shown here in purple and labeled as $S$. The existence of subsections of Datasets $P$ (orange) and $Q$ (blue) not in $S$ might indicate covariate shift. Concept shift on the other hand can be estimated for both Datasets $P$ and $Q$ for regions in $S$ (labelled $S_P$ and $S_Q$ respectively). Note that there is the same prevalence of both positive (stars) and negative (circles) examples in both $S_P$ and $S_Q$, however, their relative location has shifted for each dataset, which might indicate concept shift. *WhyShift* determines $S$ using a domain classifier and estimates the performance of an input classifier on examples in both $P$ and $Q$ and compares this to the classifier's performance on examples in $S_P$ and $S_Q$. It then used these empirical estimates of performance to estimate the proportion of performance change due to concept shift.

# 4.P Normalization aligns dataset means

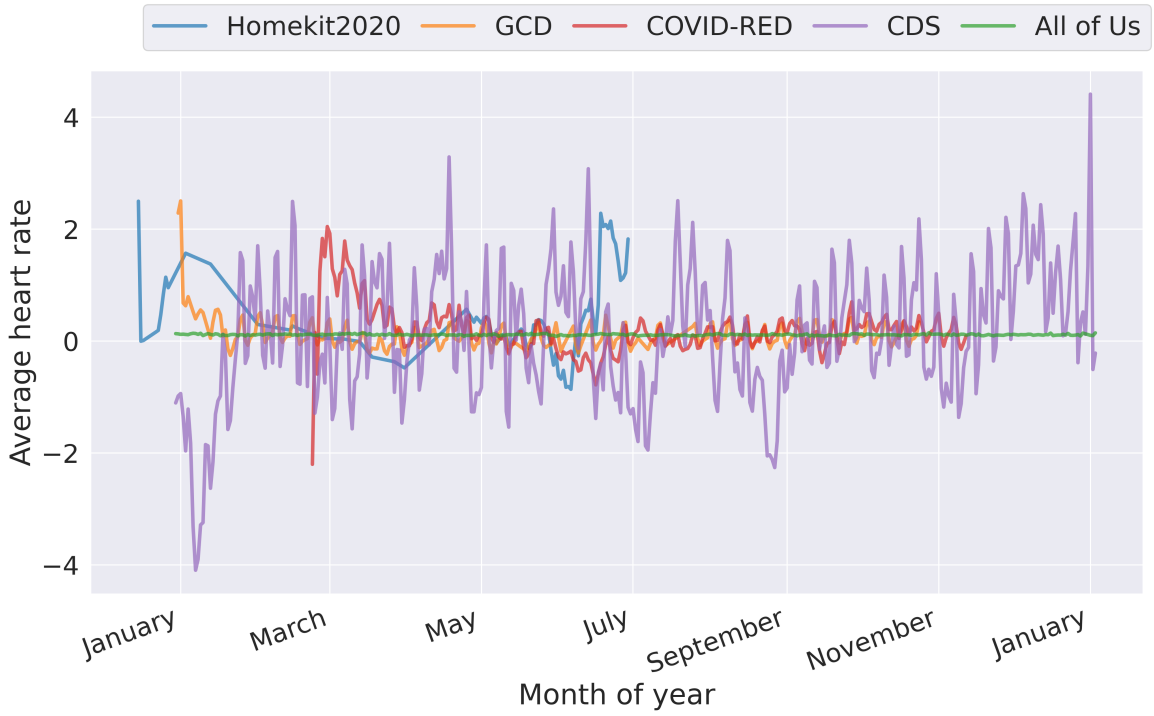


**Figure 4.P.1.** Average daily resting HR taken as the mean across all participants with available z-score normalized data (see Figure 4.M.1) on the same relative date (i.e. 2nd Tuesday of each year) and the mean across repeated relative dates for datasets spanning multiple years (*All of Us* and CDS).

## 4.17 Acknowledgments

Chapter 4, is a reprint of P. Kasl, S. Soltani, L. Keeler Bruce, V. Kumar Viswanath, W. Hartogensis, A. Gupta, I. Altintas, S. Dilchert, F.M. Hecht, A.E. Mason, and B. L. Smarr, "A Cross-study Analysis of Wearable Datasets and the Generalizability of Acute Illness Monitoring Models", 2024 currently in press at the *Conference on Health Inference and Learning*. The dissertation author was the primary author of this paper.

# Chapter 5

# Summary and future work

This thesis demonstrates the potential of wearable technologies in monitoring acute physiological changes following significant health events, notably COVID-19 vaccinations and the onset of fever. Further, I outline the feasibility of detecting fever at scale using these data. I present a comprehensive analysis of multiple longitudinal wearable device datasets, assess the generalizability of algorithms for acute illness monitoring, and characterize biases linked to demographic variables. This work demonstrates the capabilities of wearables in detecting acute physiological changes, highlighting their utility in enhancing individual and public health responses to acute physiological challenges. However, substantial challenges exist and the models currently used might be doing a poor job of capturing all the complexities of human physiology, as many models for acute illness detection struggle to make broadly generalizable predictions (Adler et al., 2022; Pillai et al., 2023; Xu et al., 2022b). Future studies should aim to refine these models; in particular, techniques that seek to derive generalizable representations from continuous wearable data might better handle differences between devices and demographic groups.

Future research in monitoring vaccine efficacy using wearable technologies could significantly enhance public health outcomes by providing real-time, individualized vaccine efficacy data. This would involve rigorous validation studies to confirm whether physiological markers measured by devices like the Oura Ring, Fitbit, and Apple Watch can reliably predict immune

responses. Subsequent work from other groups (Quer et al., 2022) demonstrated that changes of similar magnitude and directionality occur in FitBit and Apple Watch users following vaccination for COVID-19. They did not, however, determine whether these changes were correlated with eventual antibody production. Future work could extend to large-scale validation studies that examine the correlation between physiological responses detected by wearables and subsequent immune responses. It might even be possible to leverage traditional study design methods, like a double-blind, randomized, placebo-controlled trial (Kaptchuk, 2001), to gain further insight into the sensitivity of methods using wearable devices to determine vaccine effectiveness. Such studies could explore a range of vaccines and wearable devices to assess consistency and reliability across technologies. This work could then further assess the predictive power of specific physiological changes, such as changes in heart rate variability or skin temperature post-vaccination, as indicators of immune response efficacy. Additionally, exploring demographic variations in vaccine response could lead to personalized vaccination strategies, potentially adjusting vaccine dosage or schedule to optimize efficacy.

In Chapter 3 of this thesis, I make the case that wearable device data might be particularly useful for syndromic surveillance, however, there exist several barriers to implementing such a system in the real world. Systems designed to use wearable device data for surveillance at large will likely need to adapt some existing spatio-temporal methods and models from public health surveillance to wearable device data. While these methods (i.e., SaTScan (Martin Kulldorff, 2022)) have already shown substantial promise for detecting regional disease outbreaks through monitoring case-reports (Curtis et al., 2022; Greene et al., 2021), the models these methods are based on may not implicitly handle some of the nuances of wearable device data. In particular, these models often assume that the *location* of individuals is static over time. However, individuals are often quite mobile and the determining factor of a cluster of cases might be prior colocalization at a particular social gathering (Greene et al., 2021). As such, future work could attempt to incorporate historical location data as well as historical syndrome data. The model I proposed in Chapter 3 outputs the probability that a particular

day is from a fever day. This probability is a continuous variable rather, while case reports are often modeled as a Bernoulli or Poisson distribution. Methods might need to either discretize these predicted probabilities or extend these methods to a new class of "case counts". Of course, improvements in the machine learning model used to make these predictions would significantly enhance such a system. Current spatio-temporal methods might also struggle to handle the scale (i.e., number of individuals) and the temporal resolution of these data, however, these amount to computational rather than conceptual challenges. Future implementations should focus on developing frameworks that ensure the ethical handling of sensitive health data, incorporating stringent data security measures and transparent user consent processes (Chikwetu et al., 2023). Other critical considerations for such a system include ensuring user data privacy, which might include new data anonymization techniques; it is known that it is possible to reidentify individuals based on their wearable device data (Chikwetu et al., 2023). Enhancing the diversity of study populations might also improve the generalizability of these models (Arora et al., 2023). Collaboration between device manufacturers and public health agencies would also be essential to align the technological capabilities with public health needs.

Addressing the technical challenges posed by distribution shifts and label noise in datasets based on wearable devices is crucial for the advancement of this field. Future research should focus on developing machine learning models that are robust to these issues. A promising avenue for these particular datasets and tasks include semi-supervised learning (Yu and Sano, 2023), which has shown promise in other wearable-ML tasks and might enable leveraging the vast majority of unlabelled examples in these datasets; participants tend to wear their wearable devices more frequently than they complete the daily survey. Further, unsupervised domain adaptation techniques might be useful when deploying a pre-existing model in a new geographic region, with a new wearable device, or across time as it might improve performance without requiring labels in a new domain (Fernando et al., 2013; Rostami and Galstyan, 2023). Techniques from representation learning might also provide an avenue for addressing the biases along demographic or device axes (Jungo et al., 2024). Another promising area is the development of

federated learning models which allow for decentralized model fine-tuning (Wang et al., 2023). The decentralized nature of federated learning might mitigate some privacy concerns while leveraging distributed data to improve model performance. Alternatively, these datasets likely exhibit substantial levels of label noise and investigation into methods to mitigate this might prove useful (Saeed et al., 2024). They might also consider leveraging the "fuzzy" (or "soft") nature of some illness labels to improve model performance (Yuan et al., 2023). For example, the Homekit2020 symptom labels are based on a severity score (1 through 4). Preliminary results suggest that training detection models as regressors on symptom severity (instead of binary classifiers) improves within-dataset classification performance and cross-dataset generalizability. Ultimately, as researchers continue to work towards deployable models for acute infectious disease surveillance it is imperative that these technologies are effective for global disease surveillance across varying populations and geographical regions and that they have the ability to adapt to changes that will necessarily occur to their underlying data over time. Doing so will ensure their utility in the face of evolving public health challenges.

# Bibliography

(2022). OB1203 Pulse Oximeter Algorithm for SpO2, Heart Rate, and Respiration Rate.

104th Congress (1996). HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT OF 1996. https://www.govinfo.gov/content/pkg/PLAW-104publ191/html/PLAW-104publ191.htm.

Abbasi-Sureshjani, S., Raumanns, R., Michels, B. E. J., Schouten, G., and Cheplygina, V. (2020). Risk of Training Diagnostic Algorithms on Data with Demographic Bias. In Cardoso, J., Van Nguyen, H., Heller, N., Henriques Abreu, P., Isgum, I., Silva, W., Cruz, R., Pereira Amorim, J., Patel, V., Roysam, B., Zhou, K., Jiang, S., Le, N., Luu, K., Sznitman, R., Cheplygina, V., Mateus, D., Trucco, E., and Abbasi, S., editors, *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, volume 12446, pages 183–192. Springer International Publishing, Cham.

Abir, F. F., Alyafei, K., Chowdhury, M. E. H., Khandakar, A., Ahmed, R., Hossain, M. M., Mahmud, S., Rahman, A., Abbas, T. O., Zughaier, S. M., and Naji, K. K. (2022). PCovNet: A presymptomatic COVID-19 detection framework using deep learning model using wearables data. *Computers in Biology and Medicine*, 147:105682.

Adler, D. A., Wang, F., Mohr, D. C., and Choudhury, T. (2022). Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE*, 17(4):e0266516.

Alavi, A., Bogu, G. K., Wang, M., Rangan, E. S., Brooks, A. W., Wang, Q., Higgs, E., Celli, A., Mishra, T., Metwally, A. A., Cha, K., Knowles, P., Alavi, A. A., Bhasin, R., Panchamukhi, S., Celis, D., Aditya, T., Honkala, A., Rolnik, B., Hunting, E., Dagan-Rosenfeld, O., Chauhan, A., Li, J. W., Bejikian, C., Krishnan, V., McGuire, L., Li, X., Bahmani, A., and Snyder, M. P. (2022). Real-time alerting system for COVID-19 and other stress events using wearable data. *Nature Medicine*, 28(1):175–184.

Althubaiti, A. (2016). Information bias in health research: Definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*, 9:211–217.

Altini, M. and Kinnunen, H. (2021). The Promise of Sleep: A Multi-Sensor Approach for Accurate Sleep Stage Detection Using the Oura Ring. *Sensors*, 21(13):4302.

Arora, A., Alderman, J. E., Palmer, J., Ganapathi, S., Laws, E., McCradden, M. D., Oakden-Rayner, L., Pfohl, S. R., Ghassemi, M., McKay, F., Treanor, D., Rostamzadeh, N., Mateen, B., Gath, J., Adebajo, A. O., Kuku, S., Matin, R., Heller, K., Sapey, E., Sebire, N. J., Cole-Lewis, H., Calvert, M., Denniston, A., and Liu, X. (2023). The value of standards for health datasets in artificial intelligence-based applications. *Nature Medicine*, 29(11):2929–2938.

Avram, R., Tison, G. H., Aschbacher, K., Kuhar, P., Vittinghoff, E., Butzner, M., Runge, R., Wu, N., Pletcher, M. J., Marcus, G. M., and Olgin, J. (2019). Real-world heart rate norms in the Health eHeart study. *npj Digital Medicine*, 2(1):1–10.

Baden Lindsey R., El Sahly Hana M., Essink Brandon, Kotloff Karen, Frey Sharon, Novak Rick, Diemert David, Spector Stephen A., Rouphael Nadine, Creech C. Buddy, McGettigan John, Khetan Shishir, Segall Nathan, Solis Joel, Brosz Adam, Fierro Carlos, Schwartz Howard, Neuzil Kathleen, Corey Lawrence, Gilbert Peter, Janes Holly, Follmann Dean, Marovich Mary, Mascola John, Polakowski Laura, Ledgerwood Julie, Graham Barney S., Bennett Hamilton, Pajon Rolando, Knightly Conor, Leav Brett, Deng Weiping, Zhou Honghong, Han Shu, Ivarsson Melanie, Miller Jacqueline, and Zaks Tal (2021). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *New England Journal of Medicine*, 384(5):403–416.

Benedict, C. and Cedernaes, J. (2021). Could a good night's sleep improve COVID-19 vaccine efficacy? *The Lancet Respiratory Medicine*, 9(5):447–448.

Bogu, G. K. and Snyder, M. P. (2021). Deep learning-based detection of COVID-19 using wearables data. *medRxiv*, page 2021.01.08.21249474.

Brakenhoff, T. B., Goodale, B. M., Willigen, M. V., Markovic, A., Kovacevic, V., Veen, D., Mitratza, M., van de Wijgert, J., Franks, B., Montes, S., Fredslund, E. K., Korkmaz, S., Rispens, T., Risch, L., Dowling, A. V., Folarin, A. A., Bruijning, P., Dobson, R., Heikamp, T., Klaver, P., Bai, X., Grossman, K., Ornella, W., Klaver, P., Cronin, M., Grobbee, D. E., and Consortium, O. b. o. C.-R. (2023). Remote Early Detection of SARS-CoV-2 infections (COVID-RED).

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brittain, H. K., Scott, R., and Thomas, E. (2017). The rise of the genome and personalised medicine. *Clinical Medicine*, 17(6):545–551.

Cai, T. T., Namkoong, H., and Yadlowsky, S. (2023). Diagnosing Model Performance Under Distribution Shift.

Cao, R., Azimi, I., Sarhaddi, F., Niela-Vilen, H., Axelin, A., Liljeberg, P., and Rahmani, A. M. (2022). Accuracy Assessment of Oura Ring Nocturnal Heart Rate and Heart Rate Variability in Comparison With Electrocardiography in Time and Frequency Domains: Comprehensive Analysis. *Journal of Medical Internet Research*, 24(1):e27487.

Casilari, E. and Silva, C. A. (2022). An analytical comparison of datasets of Real-World and simulated falls intended for the evaluation of wearable fall alerting systems. *Measurement*, 202:111843.

CDC (2023). What to Expect after Getting a COVID-19 Vaccine.

https://archive.cdc.gov/coronavirus/2019-ncov/vaccines/expect/after.html.

CDC (2024). What to Expect at Your Appointment to Get Vaccinated for COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/expect.html.

Chandrasekaran, R., Katthula, V., and Moustakas, E. (2020). Patterns of Use and Key Predictors for the Use of Wearable Health Care Devices by US Adults: Insights from a National Survey. *Journal of Medical Internet Research*, 22(10):e22443.

Chaudhury, S., Yu, C., Liu, R., Kumar, K., Hornby, S., Duplessis, C., Sklar, J. M., Epstein, J. E., and Reifman, J. (2022). Wearables Detect Malaria Early in a Controlled Human-Infection Study. *IEEE Transactions on Biomedical Engineering*, 69(6):2119–2129.

Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., and Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science*, 383(6679):164–167.

Chikwetu, L., Miao, Y., Woldetensae, M. K., Bell, D., Goldenholz, D. M., and Dunn, J. (2023). Does deidentification of data from wearable devices give us a false sense of security? A systematic review. *The Lancet Digital Health*, 5(4):e239–e247.

Cho, P. J., Yi, J., Ho, E., Shandhi, M. M. H., Dinh, Y., Patil, A., Martin, L., Singh, G., Bent, B., Ginsburg, G., Smuck, M., Woods, C., Shaw, R., and Dunn, J. (2022). Demographic Imbalances Resulting From the Bring-Your-Own-Device Study Design. *JMIR mHealth and uHealth*, 10(4):e29510.

Cleary, J. L., Fang, Y., Sen, S., and Wu, Z. (2022). A caveat to using wearable sensor data for COVID-19 detection: The role of behavioral change after receipt of test results. *PLOS ONE*, 17(12):e0277350.

Cohen, S. N., Foster, J., Foster, P., Lou, H., Lyons, T., Morley, S., Morrill, J., Ni, H., Palmer, E., Wang, B., Wu, Y., Yang, L., and Yang, W. (2024). Subtle variation in sepsis-III definitions markedly influences predictive performance within and across methods. *Scientific Reports*, 14(1):1920.

Colón-González, F. J., Lake, I. R., Morbey, R. A., Elliot, A. J., Pebody, R., and Smith, G. E. (2018). A methodological framework for the evaluation of syndromic surveillance systems: A case study of England. *BMC Public Health*, 18(1):544.

Conroy, B., Silva, I., Mehraei, G., Damiano, R., Gross, B., Salvati, E., Feng, T., Schneider, J., Olson, N., Rizzo, A. G., Curtin, C. M., Frassica, J., and McFarlane, D. C. (2022). Real-time infection prediction with wearable physiological monitoring and AI to aid military workforce readiness during COVID-19. *Scientific Reports*, 12(1):3797.

Cook, J. A. and Collins, G. S. (2015). The rise of big clinical databases. *British Journal of Surgery*, 102(2):e93–e101.

Cooper, T. M., McKinley, P. S., Seeman, T. E., Choo, T.-H., Lee, S., and Sloan, R. P. (2015). Heart Rate Variability Predicts Levels of Inflammatory Markers: Evidence for the Vagal Anti-Inflammatory Pathway. *Brain, behavior, and immunity*, 49:94–100.

Curtis, A. J., Ajayakumar, J., Curtis, J., and Brown, S. (2022). Spatial Syndromic Surveillance and COVID-19 in the U.S.: Local Cluster Mapping for Pandemic Preparedness. *International Journal of Environmental Research and Public Health*, 19(15):8931.

Darrell, T., Kloft, M., Pontil, M., Rätsch, G., and Rodner, E. (2015). Machine Learning with Interdependent and Non-identically Distributed Data (Dagstuhl Seminar 15152). Technical report, [object Object].

Davis, S., Milechin, L., Patel, T., Hernandez, M., Ciccarelli, G., Samsi, S., Hensley, L., Goff, A., Trefry, J., Johnston, S., Purcell, B., Cabrera, C., Fleischman, J., Reuther, A., Claypool, K., Rossi, F., Honko, A., Pratt, W., and Swiston, A. (2021). Detecting Pathogen Exposure During the Non-symptomatic Incubation Period Using Physiological Data: Proof of Concept in Non-human Primates. *Frontiers in Physiology*, 12:691074.

Debes, A. K., Xiao, S., Colantuoni, E., Egbert, E. R., Caturegli, P., Gadala, A., and Milstone, A. M. (2021). Association of Vaccine Type and Prior SARS-CoV-2 Infection With Symptoms and Antibody Measurements Following Vaccination Among Health Care Workers. *JAMA Internal Medicine*, 181(12):1660–1662.

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., van Hees, V. T., Trenell, M. I., Owen, C. G., Preece, S. J., Gillions, R., Sheard, S., Peakman, T., Brage, S., and Wareham, N. J. (2017). Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLOS ONE*, 12(2):e0169649.

Drolet, B. C. and Lorenzi, N. M. (2011). Translational research: Understanding the continuum from bench to bedside. *Translational Research*, 157(1):1–5.

Dunn, J., Shandhi, M. H. H., Cho, P., Roghanizad, A., Singh, K., Wang, W., Enache, O., Stern, A., Sbahi, R., Tatar, B., Fiscus, S., Khoo, Q. X., Kuo, Y., Lu, X., Hsieh, J., Kalodzitsa, A., Bahmani, A., Alavi, A., Ray, U., Snyder, M., Ginsburg, G., Pasquale, D., Woods, C., and Shaw, R. (2022). A Method for Intelligent Allocation of Diagnostic Testing by Leveraging Data from Commercial Wearable Devices: A Case Study on COVID-19. *Research Square*, pages rs.3.rs–1490524.

El-Radhi, A. S. (2019). Pathogenesis of Fever. *Clinical Manual of Fever in Children*, pages 53–68.

EUROPEAN PARLIAMENT (2016). Regulation - 2016/679 - EN - gdpr - EUR-Lex. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Feng, S., Phillips, D. J., White, T., Sayal, H., Aley, P. K., Bibi, S., Dold, C., Fuskova, M., Gilbert, S. C., Hirsch, I., Humphries, H. E., Jepson, B., Kelly, E. J., Plested, E., Shoemaker, K., Thomas, K. M., Vekemans, J., Villafana, T. L., Lambe, T., Pollard, A. J., and Voysey, M. (2021). Correlates of protection against symptomatic and asymptomatic SARS-CoV-2 infection. *Nature Medicine*, 27(11):2032–2040.

Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. (2013). Unsupervised Visual Domain Adaptation Using Subspace Alignment. In *2013 IEEE International Conference on Computer Vision*, pages 2960–2967, Sydney, Australia. IEEE.

Fitzner, J., Qasmieh, S., Mounts, A. W., Alexander, B., Besselaar, T., Briand, S., Brown, C., Clark, S., Dueger, E., Gross, D., Hauge, S., Hirve, S., Jorgensen, P., Katz, M. A., Mafi, A., Malik, M., McCarron, M., Meerhoff, T., Mori, Y., Mott, J., Olivera, M. T. D. C., Ortiz, J. R., Palekar, R., Rebelo-de-Andrade, H., Soetens, L., Yahaya, A. A., Zhang, W., and Vandemaele, K. (2018). Revision of clinical case definitions: Influenza-like illness and severe acute respiratory infection. *Bulletin of the World Health Organization*, 96(2):122–128.

Fuller, D., Colwell, E., Low, J., Orychock, K., Tobin, M. A., Simango, B., Buote, R., Van Heerden, D., Luan, H., Cullen, K., Slade, L., and Taylor, N. G. A. (2020). Reliability and Validity of Commercially Available Wearable Devices for Measuring Steps, Energy Expenditure, and Heart Rate: Systematic Review. *JMIR mHealth and uHealth*, 8(9):e18694.

Gadaleta, M., Radin, J. M., Baca-Motes, K., Ramos, E., Kheterpal, V., Topol, E. J., Steinhubl, S. R., and Quer, G. (2021). Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. *npj Digital Medicine*, 4(1):1–10.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. (2020). A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Summits on Translational Science Proceedings*, 2020:191–200.

Ghomrawi, H. M. K., O'Brien, M. K., Carter, M., Macaluso, R., Khazanchi, R., Fanton, M., DeBoer, C., Linton, S. C., Zeineddin, S., Pitt, J. B., Bouchard, M., Figueroa, A., Kwon, S., Holl, J. L., Jayaraman, A., and Abdullah, F. (2023). Applying machine learning to consumer wearable data for the early detection of complications after pediatric appendectomy. *NPJ Digital Medicine*, 6:148.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. (2018). Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA internal medicine*, 178(11):1544–1547.

Glaser, R., Kiecolt-Glaser, J. K., Bonneau, R. H., Malarkey, W., Kennedy, S., and Hughes,

J. (1992-01/1992-02). Stress-induced modulation of the immune response to recombinant hepatitis B vaccine. *Psychosomatic Medicine*, 54(1):22.

Glynn, P. and Greenland, P. (2020). Contributions of the UK biobank high impact papers in the era of precision medicine. *European Journal of Epidemiology*, 35(1):5–10.

Goergen, C. J., Tweardy, M. J., Steinhubl, S. R., Wegerich, S. W., Singh, K., Mieloszyk, R. J., and Dunn, J. (2022). Detection and Monitoring of Viral Infections via Wearable Devices and Biometric Data. *Annual Review of Biomedical Engineering*, 24(1):null.

Golbus, J. R., Pescatore, N. A., Nallamothu, B. K., Shah, N., and Kheterpal, S. (2021). Wearable device signals and home blood pressure data across age, sex, race, ethnicity, and clinical phenotypes in the Michigan Predictive Activity & Clinical Trajectories in Health (MIPACT) study: A prospective, community-based observational study. *The Lancet Digital Health*, 3(11):e707–e715.

Goldstein, N., Eisenkraft, A., Arguello, C. J., Yang, G. J., Sand, E., Ishay, A. B., Merin, R., Fons, M., Littman, R., Nachman, D., and Gepner, Y. (2021). Exploring Early Pre-Symptomatic Detection of Influenza Using Continuous Monitoring of Advanced Physiological Parameters during a Randomized Controlled Trial. *Journal of Clinical Medicine*, 10(21):5202.

Gordan, R., Gwathmey, J. K., and Xie, L.-H. (2015). Autonomic and endocrine control of cardiovascular function. *World Journal of Cardiology*, 7(4):204–214.

Greene, S. K., Peterson, E. R., Balan, D., Jones, L., Culp, G. M., Fine, A. D., and Kulldorff, M. (2021). Detecting COVID-19 Clusters at High Spatiotemporal Resolution, New York City, New York, USA, June–July 2020 - Volume 27, Number 5—May 2021 - Emerging Infectious Diseases journal - CDC.

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., Turner, R. B., Chiu, C., Doraiswamy, P. M., Hero, A., Henao, R., Ginsburg, G. S., and Dunn, J. (2021). 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.

Hajduczok, A. G., DiJoseph, K. M., Bent, B., Thorp, A. K., Mullholand, J. B., MacKay, S. A., Barik, S., Coleman, J. J., Paules, C. I., and Tinsley, A. (2021). Physiologic Response to the Pfizer-BioNTech COVID-19 Vaccine Measured Using Wearable Devices: Prospective Observational Study. *JMIR Formative Research*, 5(8):e28568.

Hammer, G. P., du Prel, J.-B., and Blettner, M. (2009). Avoiding Bias in Observational Studies. *Deutsches Ärzteblatt International*, 106(41):664–668.

Hannoodee, S. and Nasuruddin, D. N. (2024). Acute Inflammatory Response. In *StatPearls*.

StatPearls Publishing, Treasure Island (FL).

Hassantabar, S., Stefano, N., Ghanakota, V., Ferrari, A., Nicola, G. N., Bruno, R., Marino, I. R., Hamidouche, K., and Jha, N. K. (2020). CovidDeep: SARS-CoV-2/COVID-19 Test Based on Wearable Medical Sensors and Efficient Neural Networks.

Heal, C., Harvey, A., Brown, S., Rowland, A. G., and Roland, D. (2022). The association between temperature, heart rate, and respiratory rate in children aged under 16 years attending urgent and emergency care settings. *European Journal of Emergency Medicine*, 29(6):413–416.

Heath Paul T., Galiza Eva P., Baxter David N., Boffito Marta, Browne Duncan, Burns Fiona, Chadwick David R., Clark Rebecca, Cosgrove Catherine, Galloway James, Goodman Anna L., Heer Amardeep, Higham Andrew, Iyengar Shalini, Jamal Arham, Jeanes Christopher, Kalra Philip A., Kyriakidou Christina, McAuley Daniel F., Meyrick Agnieszka, Minassian Angela M., Minton Jane, Moore Patrick, Munsoor Imrozia, Nicholls Helen, Osanlou Orod, Packham Jonathan, Pretswell Carol H., San Francisco Ramos Alberto, Saralaya Dinesh, Sheridan Ray P., Smith Richard, Soiza Roy L., Swift Pauline A., Thomson Emma C., Turner Jeremy, Viljoen Marianne E., Albert Gary, Cho Iksung, Dubovsky Filip, Glenn Greg, Rivers Joy, Robertson Andreana, Smith Kathy, and Toback Seth (2021). Safety and Efficacy of NVX-CoV2373 Covid-19 Vaccine. *New England Journal of Medicine*, 385(13):1172–1183.

Henning, K. (2004). Overview of Syndromic Surveillance What is Syndromic Surveillance? https://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm.

Hiller, K. M., Stoneking, L., Min, A., and Rhodes, S. M. (2013). Syndromic Surveillance for Influenza in the Emergency Department–A Systematic Review. *PLOS ONE*, 8(9):e73832.

Hinz, B., Cheremina, O., and Brune, K. (2008). Acetaminophen (paracetamol) is a selective cyclooxygenase-2 inhibitor in man. *The FASEB Journal*, 22(2):383–390.

Hirten, R. P., Danieletto, M., Tomalin, L., Choi, K. H., Zweig, M., Golden, E., Kaur, S., Helmus, D., Biello, A., Pyzik, R., Charney, A., Miotto, R., Glicksberg, B. S., Levin, M., Nabeel, I., Aberg, J., Reich, D., Charney, D., Bottinger, E. P., Keefer, L., Suarez-Farinas, M., Nadkarni, G. N., and Fayad, Z. A. (2021). Use of Physiological Data From a Wearable Device to Identify SARS-CoV-2 Infection and Symptoms and Predict COVID-19 Diagnosis: Observational Study. *Journal of Medical Internet Research*, 23(2):e26107.

Hirten, R. P., Tomalin, L., Danieletto, M., Golden, E., Zweig, M., Kaur, S., Helmus, D., Biello, A., Pyzik, R., Bottinger, E. P., Keefer, L., Charney, D., Nadkarni, G. N., Suarez-Farinas, M., and Fayad, Z. A. (2022). Evaluation of a machine learning approach utilizing wearable data for prediction of SARS-CoV-2 infection in healthcare workers. *JAMIA Open*, 5(2):ooac041.

Houle, D., Pélabon, C., Wagner, G. P., and Hansen, T. F. (2011). Measurement and Meaning in Biology. *The Quarterly Review of Biology*, 86(1):3–34.

Huhn, S., Axt, M., Gunga, H.-C., Maggioni, M. A., Munga, S., Obor, D., Sié, A., Boudo, V., Bunker, A., Sauerborn, R., Bärnighausen, T., and Barteit, S. (2022). The Impact of Wearable Technologies in Health Research: Scoping Review. *JMIR mHealth and uHealth*, 10(1):e34384.

Ipeirotis, P. G., Provost, F., and Wang, J. (2010). Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, New York, NY, USA. Association for Computing Machinery.

Jackson, F. L. C. (1992). Race and Ethnicity as Biological Constructs. *Ethnicity & Disease*, 2(2):120–125.

Jagannath, B., Lin, K.-C., Pali, M., Sankhala, D., Muthukumar, S., and Prasad, S. (2021). Temporal profiling of cytokines in passively expressed sweat for detection of infection using wearable device. *Bioengineering & Translational Medicine*, 6(3):e10220.

Johnson, A. E. W., Pollard, T. J., and Naumann, T. (2018). Generalizability of predictive models for intensive care unit patients.

Jungo, J., Xiang, Y., Gashi, S., and Holz, C. (2024). Representation Learning for Wearable-Based Applications in the Case of Missing Data.

Kaptchuk, T. J. (2001). The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology*, 54(6):541–549.

Karimi, D., Dou, H., Warfield, S. K., and Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759.

Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y. H., and Koo, B.-H. (2018). Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature. *Psychiatry Investigation*, 15(3):235–245.

Kinnunen, H., Rantanen, A., Kenttä, T., and Koskimäki, H. (2020). Feasible assessment of recovery and cardiovascular health: Accuracy of nocturnal HR and HRV assessed via ring PPG in comparison to medical grade ECG. *Physiological Measurement*, 41(4):04NT01.

Koch, B., Denton, E., Hanna, A., and Foster, J. G. (2021). Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research.

Kolbeinsson, A., Gade, P., Kainkaryam, R., Jankovic, F., and Foschini, L. (2021). Self-supervision of wearable sensors time-series data for influenza detection. *arXiv:2112.13755 [cs]*.

Konty, K. J., Bradshaw, B., Ramirez, E., Lee, W.-N., Signorini, A., and Foschini, L. (2019).

Influenza Surveillance Using Wearable Mobile Health Devices. *Online Journal of Public Health Informatics*, 11(1).

Krauchi, K. and Wirz-Justice, A. (1994). Circadian rhythm of heat production, heart rate, and skin and core temperature under unmasking conditions in men. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 267(3):R819–R829.

Kroenke, K. and Spitzer, R. L. (March/April 1998). Gender Differences in the Reporting of Physical and Somatoform Symptoms. *Psychosomatic Medicine*, 60(2):150.

Kryder, C. (2020a). How Accurate Is My Oura Temperature Data? https://ouraring.com/blog/temperature-validated-accurate/.

Kryder, C. (2020b). How Accurate Is Oura's Respiratory Rate? https://ouraring.com/blog/how-accurate-is-ouras-respiratory-rate/.

Kumar, M. R. S., Nayagi, D. S., G, K., and S, S. (2023). A Framework for Detection and Monitoring of COVID-19 using IoT Environment in Pre-Pandemic Life. *International Journal of Computing and Digital Systems*.

Lakshmi, B. N. and Robinson Joel, M. (2023). IoT based Illness Prediction System using Machine Learning. In *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1087–1091.

Lam, B., Catt, M., Cassidy, S., Bacardit, J., Darke, P., Butterfield, S., Alshabrawy, O., Trenell, M., and Missier, P. (2021). Using Wearable Activity Trackers to Predict Type 2 Diabetes: Machine Learning–Based Cross-sectional Study of the UK Biobank Accelerometer Cohort. *JMIR Diabetes*, 6(1):e23364.

Lammers-van der Holst, H. M., Lammers, G. J., van der Horst, G. T. J., Chaves, I., de Vries, R. D., GeurtsvanKessel, C. H., Koch, B., and van der Kuy, H. M. (2022). Understanding the association between sleep, shift work and COVID-19 vaccine immune response efficacy: Protocol of the S-CORE study. *Journal of Sleep Research*, 31(2):e13496.

Lange, T., Perras, B., Fehm, H. L., and Born, J. (2003-09/2003-10). Sleep Enhances the Human Antibody Response to Hepatitis A Vaccination. *Psychosomatic Medicine*, 65(5):831.

Larimer, K., Wegerich, S., Splan, J., Chestek, D., Prendergast, H., and Vanden Hoek, T. (2021). Personalized Analytics and a Wearable Biosensor Platform for Early Detection of COVID-19 Decompensation (DeCODe): Protocol for the Development of the COVID-19 Decompensation Index. *JMIR Research Protocols*, 10(5):e27271.

Lee, S.-M., Cho, H., and Yoon, S. M. (2017). Statistical noise reduction for robust human activity recognition. In *2017 IEEE International Conference on Multisensor Fusion and Integration*

*for Intelligent Systems (MFI)*, pages 284–288.

Lei, W., Zanchettin, C., Ho, Z. E., and Nunes Amaral, L. A. (2023). Quantifying the impact of uninformative features on the performance of supervised classification and dimensionality reduction algorithms. *APL Machine Learning*, 1(4):046118.

Li, H., Wang, Y., Wan, R., Wang, S., Li, T.-Q., and Kot, A. C. (2020). Domain Generalization for Medical Imaging Classification with Linear-Dependency Regularization.

Li, J., Zhu, H., Li, J., Wang, H., Wang, B., Luo, W., and Pan, Y. (2023). A Wearable Multi-Segment Upper Limb Tremor Assessment System for Differential Diagnosis of Parkinson's Disease Versus Essential Tremor. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3397–3406.

Liu, J., Wang, T., Cui, P., and Namkoong, H. (2023). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. *37th Conference on Neural Information Processing Systems*.

Liu, K., Ballew, C., Jacobs, D. R., Sidney, S., Savage, P. J., Dyer, A., Hughes, G., and Blanton, M. M. (1989). Ethnic differences in blood pressure, pulse rate, and related characteristics in young adults. The CARDIA study. *Hypertension*, 14(2):218–226.

Lonini, L., Shawen, N., Botonis, O., Fanton, M., Jayaraman, C., Mummidisetty, C. K., Shin, S. Y., Rushin, C., Jenz, S., Xu, S., Rogers, J. A., and Jayaraman, A. (2021). Rapid Screening of Physiological Changes Associated With COVID-19 Using Soft-Wearables and Structured Activities: A Pilot Study. *IEEE Journal of Translational Engineering in Health and Medicine*, 9:1–11.

Low, C. A., Dey, A. K., Ferreira, D., Kamarck, T., Sun, W., Bae, S., and Doryab, A. (2017). Estimation of Symptom Severity During Chemotherapy From Passively Sensed Data: Exploratory Study. *Journal of Medical Internet Research*, 19(12):e9046.

Madej, M. P., Töpfer, E., Boraschi, D., and Italiani, P. (2017). Different Regulation of Interleukin-1 Production and Activity in Monocytes and Macrophages: Innate Memory as an Endogenous Mechanism of IL-1 Inhibition. *Frontiers in Pharmacology*, 8:335.

Madigan, D., Ryan, P. B., Schuemie, M., Stang, P. E., Overhage, J. M., Hartzema, A. G., Suchard, M. A., DuMouchel, W., and Berlin, J. A. (2013). Evaluating the Impact of Database Heterogeneity on Observational Study Results. *American Journal of Epidemiology*, 178(4):645–651.

Madison, A. A., Shrout, M. R., Renna, M. E., and Kiecolt-Glaser, J. K. (2021). Psychological and Behavioral Predictors of Vaccine Efficacy: Considerations for COVID-19. *Perspectives on Psychological Science*, 16(2):191–203.

Maeda, Y., Sekine, M., and Tamura, T. (2011). The Advantages of Wearable Green Reflected Photoplethysmography. *Journal of Medical Systems*, 35(5):829–834.

Mandl, K. D., Overhage, J. M., Wagner, M. M., Lober, W. B., Sebastiani, P., Mostashari, F., Pavlin, J. A., Gesteland, P. H., Treadwell, T., Koski, E., Hutwagner, L., Buckeridge, D. L., Aller, R. D., and Grannis, S. (2004). Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience. *Journal of the American Medical Informatics Association : JAMIA*, 11(2):141–150.

Mangalam, M., Sadri, A., Hayano, J., Watanabe, E., Kiyono, K., and Kelty-Stephen, D. G. (2023). Reproducible biomarkers: Leveraging nonlinear descriptors in the face of non-ergodicity.

Martin Kulldorff (2022). SaTScan - Software for the spatial, temporal, and space-time scan statistics.

Mason, A. E., Hecht, F. M., Davis, S. K., Natale, J. L., Hartogensis, W., Damaso, N., Claypool, K. T., Dilchert, S., Dasgupta, S., Purawat, S., Viswanath, V. K., Klein, A., Chowdhary, A., Fisher, S. M., Anglo, C., Puldon, K. Y., Veasna, D., Prather, J. G., Pandya, L. S., Fox, L. M., Busch, M., Giordano, C., Mercado, B. K., Song, J., Jaimes, R., Baum, B. S., Telfer, B. A., Philipson, C. W., Collins, P. P., Rao, A. A., Wang, E. J., Bandi, R. H., Choe, B. J., Epel, E. S., Epstein, S. K., Krasnoff, J. B., Lee, M. B., Lee, S.-W., Lopez, G. M., Mehta, A., Melville, L. D., Moon, T. S., Mujica-Parodi, L. R., Noel, K. M., Orosco, M. A., Rideout, J. M., Robishaw, J. D., Rodriguez, R. M., Shah, K. H., Siegal, J. H., Gupta, A., Altintas, I., and Smarr, B. L. (2022). Detection of COVID-19 using multimodal data from a wearable device: Results from the first TemPredict Study. *Scientific Reports*, 12(1):3463.

Mason, A. E., Kasl, P., Soltani, S., Green, A., Hartogensis, W., Dilchert, S., Chowdhary, A., Pandya, L. S., Siwik, C. J., Foster, S. L., Nyer, M., Lowry, C. A., Raison, C. L., Hecht, F. M., and Smarr, B. L. (2024). Elevated body temperature is associated with depressive symptoms: Results from the TemPredict Study. *Scientific Reports*, 14(1):1884.

Master, H., Annis, J., Huang, S., Beckman, J. A., Ratsimbazafy, F., Marginean, K., Carroll, R., Natarajan, K., Harrell, F. E., Roden, D. M., Harris, P., and Brittain, E. L. (2022). Association of step counts over time with the risk of chronic disease in the All of Us Research Program. *Nature Medicine*, 28(11):2301–2308.

Mayer, C., Tyler, J., Fang, Y., Flora, C., Frank, E., Tewari, M., Choi, S. W., Sen, S., and Forger, D. B. (2022). Consumer-grade wearables identify changes in multiple physiological systems during COVID-19 disease progression. *Cell Reports Medicine*, 3(4):100601.

McCoy, T., Pavlick, E., and Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Merrill, M. A. and Althoff, T. (2022). Self-supervised Pretraining and Transfer Learning Enable Flu and COVID-19 Predictions in Small Mobile Sensing Datasets.

Merrill, M. A., Safranchik, E., Kolbeinsson, A., Gade, P., Ramirez, E., Schmidt, L., Foschini, L., and Althoff, T. (2023). Homekit2020: A Benchmark for Time Series Classification on a Large Mobile Sensing Dataset with Laboratory Tested Ground Truth of Influenza Infections. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 207–228. PMLR.

Mezlini, A., Shapiro, A., Daza, E. J., Caddigan, E., Ramirez, E., Althoff, T., and Foschini, L. (2022). Estimating the Burden of Influenza-like Illness on Daily Activity at the Population Scale Using Commercial Wearable Sensors. *JAMA Network Open*, 5(5):e2211958.

Miller, D. J., Capodilupo, J. V., Lastella, M., Sargent, C., Roach, G. D., Lee, V. H., and Capodilupo, E. R. (2020). Analyzing changes in respiratory rate to predict the risk of COVID-19 infection. *PLOS ONE*, 15(12):e0243693.

Mishra, T., Wang, M., Metwally, A. A., Bogu, G. K., Brooks, A. W., Bahmani, A., Alavi, A., Celli, A., Higgs, E., Dagan-Rosenfeld, O., Fay, B., Kirkpatrick, S., Kellogg, R., Gibson, M., Wang, T., Hunting, E. M., Mamic, P., Ganz, A. B., Rolnik, B., Li, X., and Snyder, M. P. (2020). Pre-symptomatic detection of COVID-19 from smartwatch data. *Nature Biomedical Engineering*, 4(12):1208–1220.

Mitratza, M., Goodale, B. M., Shagadatova, A., Kovacevic, V., van de Wijgert, J., Brakenhoff, T. B., Dobson, R., Franks, B., Veen, D., Folarin, A. A., Stolk, P., Grobbee, D. E., Cronin, M., and Downward, G. S. (2022). The performance of wearable sensors in the detection of SARS-CoV-2 infection: A systematic review. *The Lancet. Digital Health*, 4(5):e370–e383.

Miyawaki, M., Brahim, W., Iida, Y., and Ma, J. (2023). Recognition of Psychological Stress Levels Using Wearable Biosensors. *International Symposium on Affective Science and Engineering*, ISASE2023:1–4.

Mohd-Yasin, F., Korman, C. E., and Nagel, D. J. (2003). Measurement of noise characteristics of MEMS accelerometers. *Solid-State Electronics*, 47(2):357–360.

Morales-Núñez, J. J., Muñoz-Valle, J. F., Meza-López, C., Wang, L.-F., Machado Sulbarán, A. C., Torres-Hernández, P. C., Bedolla-Barajas, M., De la O-Gómez, B., Balcázar-Félix, P., and Hernández-Bello, J. (2021). Neutralizing Antibodies Titers and Side Effects in Response to BNT162b2 Vaccine in Healthcare Workers with and without Prior SARS-CoV-2 Infection. *Vaccines*, 9(7):742.

Nagaraj, S., Gerych, W., Tonekaboni, S., Goldenberg, A., Ustun, B., and Hartvigsen, T. (2024). Learning from Time Series under Temporal Label Noise.

Nair, N. R., Schmid, L., Rueda, F. M., Pauly, M., Fink, G. A., and Reining, C. (2023). Dataset

Bias in Human Activity Recognition.

Natarajan, A., Su, H.-W., and Heneghan, C. (2020). Assessment of physiological signs associated with COVID-19 measured using wearable devices. *npj Digital Medicine*, 3(1):1–8.

Nelson, L. R., Nelson, L. A., and Zaichkowsky, L. D. (1979). A Case for Using Multiple Regression Instead of ANOVA in Educational Research. *The Journal of Experimental Education*, 47(4):324–330.

Nestor, B., Hunter, J., Kainkaryam, R., Drysdale, E., Inglis, J. B., Shapiro, A., Nagaraj, S., Ghassemi, M., Foschini, L., and Goldenberg, A. (2021). Dear Watch, Should I Get a COVID-19 Test? Designing deployable machine learning for wearables.

Nestor, B., Hunter, J., Kainkaryam, R., Drysdale, E., Inglis, J. B., Shapiro, A., Nagaraj, S., Ghassemi, M., Foschini, L., and Goldenberg, A. (2023). Machine learning COVID-19 detection from wearables. *The Lancet Digital Health*, 5(4):e182–e184.

Nobre, F. F., Monteiro, A. B. S., Telles, P. R., and Williamson, G. D. (2001). Dynamic linear model and SARIMA: A comparison of their forecasting performance in epidemiology. *Statistics in Medicine*, 20(20):3051–3069.

Parihar, A., Eubank, T. D., and Doseff, A. I. (2010). Monocytes and Macrophages Regulate Immunity through Dynamic Networks of Survival and Cell Death. *Journal of Innate Immunity*, 2(3):204–215.

Pavlov, V. A. and Tracey, K. J. (2012). The vagus nerve and the inflammatory reflex—linking immunity and metabolism. *Nature reviews. Endocrinology*, 8(12):743–754.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Pho, G. N., Thigpen, N., Patel, S., and Tily, H. (2023). Feasibility of Measuring Physiological Responses to Breakthrough Infections and COVID-19 Vaccine Using a Wearable Ring Sensor. *Digital Biomarkers*, 7(1):1–6.

Pillai, A., Nepal, S. K., Wang, W., Nemesure, M., Heinz, M., Price, G., Lekkas, D., Collins, A. C., Griffin, T., Buck, B., Preum, S. M., Cohen, T., Jacobson, N. C., Ben-Zeev, D., and Campbell, A. (2023). Investigating Generalizability of Speech-based Suicidal Ideation Detection Using Mobile Phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):1–38.

Pivovarov, R., Perotte, A. J., Grave, E., Angiolillo, J., Wiggins, C. H., and Elhadad, N. (2015).

Learning Probabilistic Phenotypes from Heterogeneous EHR Data. *Journal of biomedical informatics*, 58:156–165.

Pouwels, K. B., Pritchard, E., Matthews, P. C., Stoesser, N., Eyre, D. W., Vihta, K.-D., House, T., Hay, J., Bell, J. I., Newton, J. N., Farrar, J., Crook, D., Cook, D., Rourke, E., Studley, R., Peto, T. E. A., Diamond, I., and Walker, A. S. (2021). Effect of Delta variant on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK. *Nature Medicine*, 27(12):2127–2135.

Prather, A. A., Pressman, S. D., Miller, G. E., and Cohen, S. (2021). Temporal Links Between Self-Reported Sleep and Antibody Responses to the Influenza Vaccine. *International Journal of Behavioral Medicine*, 28(1):151–158.

Prymula, R., Siegrist, C.-A., Chlibek, R., Zemlickova, H., Vackova, M., Smetana, J., Lommel, P., Kaliskova, E., Borys, D., and Schuerman, L. (2009). Effect of prophylactic paracetamol administration at time of vaccination on febrile reactions and antibody responses in children: Two open-label, randomised controlled trials. *The Lancet*, 374(9698):1339–1350.

Quer, G., Gadaleta, M., Radin, J. M., Andersen, K. G., Baca-Motes, K., Ramos, E., Topol, E. J., and Steinhubl, S. R. (2022). Inter-individual variation in objective measure of reactogenicity following COVID-19 vaccination via smartwatches and fitness bands. *npj Digital Medicine*, 5(1):1–9.

Quer, G., Radin, J. M., Gadaleta, M., Baca-Motes, K., Ariniello, L., Ramos, E., Kheterpal, V., Topol, E. J., and Steinhubl, S. R. (2021). Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nature Medicine*, 27(1):73–77.

Radin, J. M., Wineinger, N. E., Topol, E. J., and 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.

Rao, C., Di Lascio, E., Demanse, D., Marshall, N., Sopala, M., and De Luca, V. (2023). Association of digital measures and self-reported fatigue: A remote observational study in healthy participants and participants with chronic inflammatory rheumatic disease. *Frontiers in Digital Health*, 5.

Realdi, G., Previato, L., and Vitturi, N. (2008). Selection of diagnostic tests for clinical decision making and translation to a problem oriented medical record. *Clinica Chimica Acta*, 393(1):37–43.

Richards, D. M., Tweardy, M. J., Steinhubl, S. R., Chestek, D. W., Hoek, T. L. V., Larimer, K. A., and Wegerich, S. W. (2021). Wearable sensor derived decompensation index for continuous remote monitoring of COVID-19 diagnosed patients. *npj Digital Medicine*, 4(1):1–11.

Riester, E., Findeisen, P., Hegel, J. K., Kabesch, M., Ambrosch, A., Rank, C. M., Pessl, F., Laengin, T., and Niederhauser, C. (2021). Performance evaluation of the Roche Elecsys Anti-SARS-CoV-2 S immunoassay. *Journal of Virological Methods*, 297:114271.

Risch, M., Grossmann, K., Aeschbacher, S., Weideli, O. C., Kovac, M., Pereira, F., Wohlwend, N., Risch, C., Hillmann, D., Lung, T., Renz, H., Twerenbold, R., Rothenbühler, M., Leibovitz, D., Kovacevic, V., Markovic, A., Klaver, P., Brakenhoff, T. B., Franks, B., Mitratza, M., Downward, G. S., Dowling, A., Montes, S., Grobbee, D. E., Cronin, M., Conen, D., Goodale, B. M., and Risch, L. (2022). Investigation of the use of a sensor bracelet for the presymptomatic detection of changes in physiological parameters related to COVID-19: An interim analysis of a prospective cohort study (COVI-GAPP). *BMJ Open*, 12(6):e058274.

Roche Diagnostics (2022). Elecsys Anti-SARS-CoV-2 S.

Roelofs, R. (2019). Measuring Generalization and Overfitting in Machine Learning. *UC Berkeley*.

Rostami, M. and Galstyan, A. (2023). Overcoming Concept Shift in Domain-Aware Settings through Consolidated Internal Distributions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9623–9631.

Roulston, M. S. (2007). Performance targets and the Brier score. *Meteorological Applications*, 14(2):185–194.

Saeed, A., Spathis, D., Oh, J., Choi, E., and Etemad, A. (2024). Learning under Label Noise through Few-Shot Human-in-the-Loop Refinement.

Sahoo, R., Zhao, S., Chen, A., and Ermon, S. (2021). Reliable Decisions with Threshold Calibration. In *Advances in Neural Information Processing Systems*, volume 34, pages 1831–1844. Curran Associates, Inc.

Saleh, E., Moody, M. A., and Walter, E. B. (2016). Effect of antipyretic analgesics on immune responses to vaccination. *Human Vaccines & Immunotherapeutics*, 12(9):2391–2402.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3):160.

Schoeler, T., Speed, D., Porcu, E., Pirastu, N., Pingault, J.-B., and Kutalik, Z. (2023). Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour*, 7(7):1216–1227.

Shapiro, A., Marinsek, N., Clay, I., Bradshaw, B., Ramirez, E., Min, J., Trister, A., Wang, Y., Althoff, T., and Foschini, L. (2021). Characterizing COVID-19 and Influenza Illnesses in the Real World via Person-Generated Health Data. *Patterns*, 2(1):100188.

Shcherbina, A., Mattsson, C. M., Waggott, D., Salisbury, H., Christle, J. W., Hastie, T., Wheeler, M. T., and Ashley, E. A. (2017). Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine*, 7(2):3.

Shiba, S. K., Temple, C. A., Krasnoff, J., Dilchert, S., Smarr, B. L., Robishaw, J., Mason, A. E., Shiba, S. K., Temple, C. A., Krasnoff, J., Dilchert, S., Smarr, B. L., Robishaw, J., and Mason, A. E. (2023). Assessing Adherence to Multi-Modal Oura Ring Wearables From COVID-19 Detection Among Healthcare Workers. *Cureus*, 15(9).

Shih, C.-H., Lin, Y.-J., Lee, K.-F., Chien, P.-Y., and Drake, P. (2010). Real-time electronic nose based pathogen detection for respiratory intensive care patients. *Sensors and Actuators B: Chemical*, 148(1):153–157.

Shim, J., Fleisch, E., and Barata, F. (2023). Circadian Rhythm Analysis Using Wearable-Based Accelerometry as a Digital Biomarker of Aging and Healthspan. Preprint, In Review.

Shimoni, Z., Niven, M., Kama, N., Dusseldorp, N., and Froom, P. (2008). Increased complaints of fever in the emergency room can identify influenza epidemics. *European Journal of Internal Medicine*, 19(7):494–498.

Shimoni, Z., Rodrig, J., Dusseldorp, N., Niven, M., and Froom, P. (2012). Increased emergency department chief complaints of fever identified the influenza (H1N1) pandemic before outpatient symptom surveillance. *Environmental Health and Preventive Medicine*, 17(1):69–72.

Singh, A. (2021). CLDA: Contrastive Learning for Semi-Supervised Domain Adaptation. In *Advances in Neural Information Processing Systems*, volume 34, pages 5089–5101. Curran Associates, Inc.

Singh, H., Mhasawade, V., and Chunara, R. (2022). Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. *PLOS Digital Health*, 1(4):e0000023.

Skibińska, J. (2023). *Machine Learning-Aided Monitoring and Prediction of Respiratory and Neurodegenerative Diseases Using Wearables*. Omakustanne/Self-published.

Smarr, B. L., Aschbacher, K., Fisher, S. M., Chowdhary, A., Dilchert, S., Puldon, K., Rao, A., Hecht, F. M., and Mason, A. E. (2020). Feasibility of continuous fever monitoring using wearable devices. *Scientific Reports*, 10(1):21640.

Smith, G. E., Elliot, A. J., Lake, I., Edeghere, O., Morbey, R., Catchpole, M., Heymann, D. L., Hawker, J., Ibbotson, S., McCloskey, B., and Pebody, R. (2019). Syndromic surveillance: Two decades experience of sustainable systems – its people not just data! *Epidemiology and Infection*, 147:e101.

Spiegel, K., Sheridan, J. F., and Van Cauter, E. (2002). Effect of Sleep Deprivation on Response to Immunizaton. *JAMA*, 288(12):1471–1472.

Su, W., Yuan, Y., and Zhu, M. (2015). A Relationship between the Average Precision and the Area Under the ROC Curve. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 349–352, New York, NY, USA. Association for Computing Machinery.

Suriyakumar, V. M., Ghassemi, M., and Ustun, B. (2023). When Personalization Harms Performance: Reconsidering the Use of Group Attributes in Prediction. In *Proceedings of the 40th International Conference on Machine Learning*, pages 33209–33228. PMLR.

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers, J. J., and Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, 36(2):747–756.

The All of Us Research Program Investigators (2019). The "All of Us" Research Program. *The New England journal of medicine*, 381(7):668–676.

Therneau, T. M. and Grambsch, P. M. (2000). The Cox Model. In Therneau, T. M. and Grambsch, P. M., editors, *Modeling Survival Data: Extending the Cox Model*, pages 39–77. Springer, New York, NY.

Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, Colorado Springs, CO, USA. IEEE.

Uysal, E. B., Gümüş, S., Bektöre, B., Bozkurt, H., and Gözalan, A. (2022). Evaluation of antibody response after COVID-19 vaccination of healthcare workers. *Journal of Medical Virology*, 94(3):1060–1066.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. (2019). Evaluating model calibration in classification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR.

Vatner, S. F. and Pagani, M. (1976). Cardiovascular adjustments to exercise: Hemodynamics and mechanisms. *Progress in Cardiovascular Diseases*, 19(2):91–108.

Vaughn, J., Baral, A., Vadari, M., and Boag, W. (2020). Dataset Bias in Diagnostic AI systems: Guidelines for Dataset Collection and Usage. *Proceedings of CHIL '20: ACM The ACM Conference on Health, Inference, and Learning (CHIL '20)*.

Vittinghoff, E. and McCulloch, C. E. (2007). Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*, 165(6):710–718.

Vorburger, P. and Bernstein, A. (2006). Entropy-based Concept Shift Detection. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 1113–1118.

Wacholder, S., Rothman, N., and Caporaso, N. (2000). Population Stratification in Epidemiologic Studies of Common Genetic Variants and Cancer: Quantification of Bias. *JNCI: Journal of the National Cancer Institute*, 92(14):1151–1158.

Wainberg, M., Jones, S. E., Beaupre, L. M., Hill, S. L., Felsky, D., Rivas, M. A., Lim, A. S. P., Ollila, H. M., and Tripathy, S. J. (2021). Association of accelerometer-derived sleep measures with lifetime psychiatric diagnoses: A cross-sectional study of 89,205 participants from the UK Biobank. *PLOS Medicine*, 18(10):e1003782.

Wang, W., Li, X., Qiu, X., Zhang, X., Brusic, V., and Zhao, J. (2023). A privacy preserving framework for federated learning in smart healthcare systems. *Information Processing & Management*, 60(1):103167.

Weimer, W. B. (2023). Problems of Measurement and Meaning in Biology. In Weimer, W. B., editor, *Epistemology of the Human Sciences: Restoring an Evolutionary Approach to Biology, Economics, Psychology and Philosophy*, pages 53–70. Springer International Publishing, Cham.

Wells, B. J., Chagin, K. M., Nowacki, A. S., and Kattan, M. W. (2013). Strategies for Handling Missing Data in Electronic Health Record Derived Data. *eGEMs*, 1(3):1035.

Westbrook, J. I., Coiera, E., Dunsmuir, W. T. M., Brown, B. M., Kelk, N., Paoloni, R., and Tran, C. (2010). The impact of interruptions on clinical task completion. *BMJ Quality & Safety*, 19(4):284–289.

Wiedermann, M., Bruckmann, R., and Brockmann, D. (2023). Corona-Datenspende - Teildatensatz Vitaldaten.

Wilhelm, A., Widera, M., Grikscheit, K., Toptan, T., Schenk, B., Pallas, C., Metzler, M., Kohmer, N., Hoehl, S., Marschalek, R., Herrmann, E., Helfritz, F. A., Wolf, T., Goetsch, U., and Ciesek, S. (2022). Limited neutralisation of the SARS-CoV-2 Omicron subvariants BA.1 and BA.2 by convalescent and vaccine serum and monoclonal antibodies. *eBioMedicine*, 82.

Xu, X., Liu, X., Zhang, H., Wang, W., Nepal, S., Sefidgar, Y., Seo, W., Kuehn, K. S., Huckins, J. F., Morris, M. E., Nurius, P. S., Riskin, E. A., Patel, S., Althoff, T., Campbell, A., Dey, A. K., and Mankoff, J. (2022a). GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–34.

Xu, X., Zhang, H., Sefidgar, Y., Ren, Y., Liu, X., Seo, W., Brown, J., Kuehn, K., Merrill, M., Nurius, P., Patel, S., Althoff, T., Morris, M. E., Riskin, E., Mankoff, J., and Dey, A. K.

(2022b). GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. *36th Conference on Neural Information Processing Systems*.

Yamagami, K., Nomura, A., Kometani, M., Shimojima, M., Sakata, K., Usui, S., Furukawa, K., Takamura, M., Okajima, M., Watanabe, K., and Yoneda, T. (2021). Early Detection of Symptom Exacerbation in Patients With SARS-CoV-2 Infection Using the Fitbit Charge 3 (DEXTERITY): Pilot Evaluation. *JMIR Formative Research*, 5(9):e30819.

Yamagata, T., Tonkin, E. L., Sanchez, B. A., Craddock, I., Nieto, M. P., Santos-Rodriguez, R., Yang, W., and Flach, P. (2023). When the Ground Truth is not True: Modelling Human Biases in Temporal Annotations.

Yu, H. and Sano, A. (2023). Semi-Supervised Learning for Wearable-based Momentary Stress Detection in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(2):80:1–80:23.

Yuan, H., Xu, N., Shi, Y., Geng, X., and Rui, Y. (2023). Learning From Biased Soft Labels.

Zhang, L., Zhu, Y., Jiang, M., Wu, Y., Deng, K., and Ni, Q. (2021). Body Temperature Monitoring for Regular COVID-19 Prevention Based on Human Daily Activity Recognition. *Sensors (Basel, Switzerland)*, 21(22):7540.