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

Patrick Kasl PKASL@UCSD.EDU
Shu Chien-Gene Lay Department of Bioengineering, University of California San Diego, San Diego, CA, USA
Severine Soltani SSOLTANI@UCSD.EDU
Lauryn Keeler Bruce LBRUCE@UCSD.EDU
UC San Diego Health Department of Biomedical Informatics, University of California San Diego, San Diego, CA, USA
Varun Kumar Viswanath VKVISWAN@UCSD.EDU
Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA, USA
Wendy Hartogensis WENDY.HARTOGENSIS@UCSF.EDU
Osher Center for Integrative Health, University of California San Francisco, San Francisco, CA, USA
Amarnath Gupta A1GUPTA@UCSD.EDU
Ilkay Altintas IALTINTAS@UCSD.EDU
Halıcıoğlu Data Science Institute and San Diego Supercomputer Center, University of California San Diego, San Diego, CA, USA
Stephan Dilchert STEPHAN.DILCHERT@BARUCH.CUNY.EDU
Department of Management, Zicklin School of Business, Baruch College, The City University of New York, New York, NY, USA
Frederick M. Hecht RICK.HECHT@UCSF.EDU
Ashley Mason ASHLEY.MASON@UCSF.EDU
Osher Center for Integrative Health, University of California San Francisco, San Francisco, CA, USA
Benjamin L. Smarr BSMARR@UCSD.EDU
Shu Chien-Gene Lay Department of Bioengineering, University of California San Diego, San Diego, CA, USA

Abstract

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.

Data and Code Availability This paper uses five datasets from wearable devices. All but one (TemPredict) are publicly available: 1) FitBit data from the Homekit2020 study (Merrill et al., 2023), 2) Oura Ring data from the TemPredict, 3) Ava watch data from the COVID-RED study (Brakenhoff et al., 2023), 4) FitBit data from the *All of*

Us study (All, 2019), and 5) Apple Watch, Amazfit, Fitbit, Garmin, Google Fit, Huawei, Oura, Polar, Samsung Health and Withings data from the Corona-Datenspende study (Wiedermann et al., 2023). Code for processing the publicly available datasets can be found at this GitHub repository.

Institutional Review Board (IRB) The study that collected the TemPredict dataset was approved by (IRB, IRB# 20-30408) and the U.S. Department of Defense (DOD) Human Research Protections Office (HRPO, HRPO# E01877.1a) approved all study activities, and all research was performed per relevant guidelines and regulations and the Declaration of Helsinki. All participants provided informed consent. All other datasets are publicly available and do not require IRB approval.

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 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

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

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.

2. Related Work

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, 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.

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.

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

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

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.

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, TemPredict, 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, TemPredict, 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 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. Appendices B, C, D, E and F include further details on features, preprocessing steps, and details on how to access each dataset.

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 Appendix B for more details.

3.2. TemPredict Dataset

The TemPredict dataset 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 Appendix C for additional details.

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 TemPredict 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 Appendix D for additional details.

3.4. *All of Us*

The *All of Us* research program is an ongoing major initiative to collect diverse health-related data, in-

Table 1: Descriptions of the datasets used in these analyses. See Appendix I for participant counts expanded by demographics.

Dataset	Device	Number of Participants	Demographics	Features	Questionnaires
Homekit2020	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, temperature	Symptoms, COVID and flu test results
COVID-RED	Ava smartwatch	n=14,955	Sex, ethnicity, age, education, BMI, etc.	HR, HRV, RR, temperature, 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

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

cluding electronic health records, genomic data, physical measurements, participant questionnaires, and wearable device data from over a million Americans (All, 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 generalizability. See Appendix E for additional details, particularly how we calculated resting HR from minute-level data.

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 Appendix F for further details.

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?

4.1. 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 5 for the total numbers in each category.

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

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.

4.3. 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 Appendix A for criteria and Appendices 7, 8 and 9 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.

4.3.1. Ground truth definitions

Prior acute illness monitoring studies have a wide range of ground truth definitions (see Table 8 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 Appendix M 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 Appendices B, C and 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.

4.3.2. Normalization strategy

Eleven of the thirteen acute illness monitoring studies we reviewed used a lagged, within-individual z-score normalization (Appendix 7). 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 Appendix L and we found that z-scoring by a ten-day window with a twelve-day offset was optimal for these data.

4.3.3. 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, TemPredict, 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 4.3.2) prior to a ground truth day (see

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

Appendix L 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, Appendix M) and that within-dataset performance is lower when limiting features to those that are shared across datasets (HR and sleep vs all available features, Appendix M).

4.3.4. 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 9), 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 Appendix L for implementation details.

4.3.5. 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 12.

4.3.6. 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. 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 Appendix N for implementation details and a schematic further describing how *WhyShift* estimates performance changes due to concept shift.

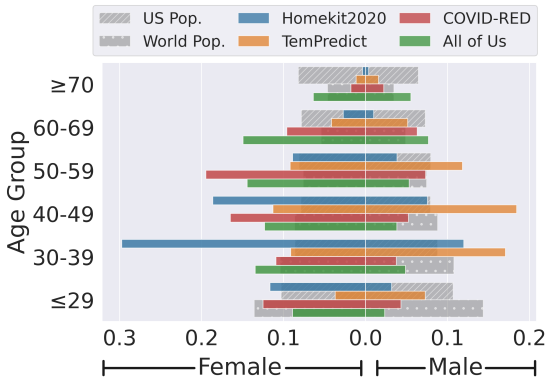

Figure 1: Datasets are biased in self-reported age and sex, relative to both the U.S. and World populations. Within-dataset participant counts are normalized by the total number of participants in each dataset and displayed using a population pyramid.

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.

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 1) as well as White groups (Figure 2).

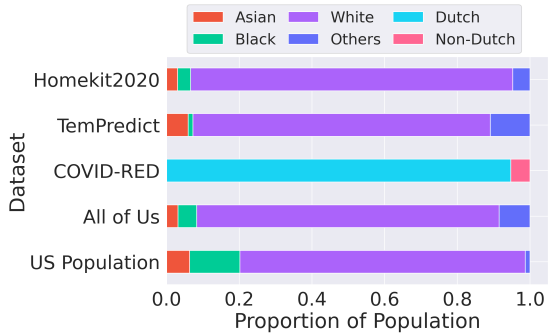

Figure 2: Datasets are biased in self-reported ethnicity as compared to the U.S. population, particularly with respect to Black participants. Within-dataset participant counts are normalized by the total number of participants in each dataset and are displayed based on the relative prevalence of self-reported ethnicity.

5.2. Average dataset resting HR

There appear to be substantial differences in the underlying distributions of within-dataset average resting HR (Figure 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, TemPredict, and *All of Us* datasets in Figures 4, 5 and 6 along with descriptions of weekday vs. weekend differences across datasets (Appendix G).

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 2) and African-American participants exhibit higher HRs relative to white participants. Notably, age is positively correlated with HR in the

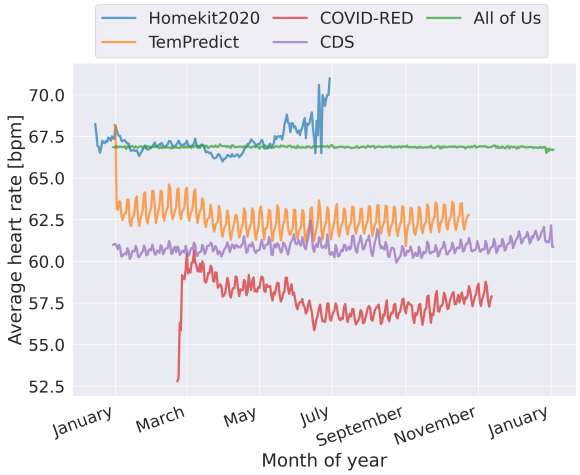

Figure 3: Within dataset mean resting HR varies substantially between datasets. Here, the average daily resting HR was taken as the mean across all participants with available data on the same relative date (i.e., 2nd Tuesday of each year) and the mean across repeated relative dates for datasets spanning multiple years (*All of Us* and CDS).

TemPredict dataset, while age is negatively correlated with HR in the *All of Us* dataset.

5.4. HR differences across datasets

Multiple regression confirms the qualitative assessment observed in Figure 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 TemPredict dataset (4.86 bpm lower) and the COVID-RED dataset (10.41 bpm lower, Table 6). 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).

5.5. Acute illness monitoring generalizability

In general, performance was worse on external datasets, however, this was not always the case (Table 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 TemPredict and COVID-RED datasets. Models trained on the TemPredict dataset (the largest dataset by number of training

Table 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)
TemPredict	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 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	TemPredict	COVID-RED	Mean others	Percent drop
Viral	Homekit2020	**0.780**	0.496	0.586	0.54	30.64
	TemPredict	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
	TemPredict	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
	TemPredict	0.679	**0.673**	0.694	0.69	-2.01
	COVID-RED	0.653	0.630	**0.685**	0.64	6.35

Table 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	TemPredict	COVID-RED
Viral	Homekit2020	-	0.91	0.94
	TemPredict	0.98	-	1.0
	COVID-RED	0.99	1.0	-
Flu	Homekit2020	-	0.77	0.55
	TemPredict	0.76	-	0.72
	COVID-RED	0.613	0.74	-
Fever	Homekit2020	-	1.0	0.77
	TemPredict	1.0	-	0.73
	COVID-RED	0.81	0.78	-

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

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). The proportion of performance change attributable to concept shift was also approximately symmetric for each task (e.g., trained on COVID-RED, tested on TemPredict was close to trained on TemPredict, 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.

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 (Appendix O). 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). 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.

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.

References

The "All of Us" Research Program. *The New England journal of medicine*, 381(7):668–676, August 2019. ISSN 0028-4793. doi: 10.1056/NEJMsr1809937.

Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt E. J. Michels, Gerard Schouten, and Veronika Cheplygina. Risk of Training Diagnostic Algorithms on Data with Demographic Bias. In Jaime Cardoso, Hien Van Nguyen, Nicholas Heller, Pedro Henriques Abreu, Ivana Isgum, Wilson Silva, Ricardo Cruz, Jose Pereira Amorim, Vishal Patel, Badri Roysam, Kevin Zhou, Steve Jiang, Ngan Le, Khoa Luu, Raphael Sznitman, Veronika Cheplygina, Diana Mateus, Emanuele Trucco, and Samaneh Abbasi, editors, *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, volume 12446, pages 183–192. Springer International Publishing, Cham, 2020. ISBN 978-3-030-61165-1 978-3-030-61166-8. doi: 10.1007/978-3-030-61166-8_20.

Farhan Fuad Abir, Khalid Alyafei, Muhammad E. H. Chowdhury, Amith Khandakar, Rashid Ahmed, Muhammad Maqsud Hossain, Sakib Mahmud, Ashiqur Rahman, Tareq O. Abbas, Susu M. Zughaier, and Khalid Kamal Naji. PCovNet: A presymptomatic COVID-19 detection framework using deep learning model using wearables data. *Computers in Biology and Medicine*, 147:105682, August 2022. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2022.105682.

Daniel A. Adler, Fei Wang, David C. Mohr, and Tanzeem Choudhury. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE*, 17(4):e0266516, April 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0266516.

Arash Alavi, Gireesh K. Bogu, Meng Wang, Ekanath Srihari Rangan, Andrew W. Brooks, Qiwen Wang, Emily Higgs, Alessandra Celli, Tejaswini Mishra, Ahmed A. Metwally, Kexin Cha, Peter Knowles, Amir A. Alavi, Rajat Bhasin, Shrinivas Panchamukhi, Diego Celis, Tagore Aditya, Alexander Honkala, Benjamin Rolnik, Erika Hunting, Orit Dagan-Rosenfeld, Arshdeep Chauhan, Jessi W. Li, Caroline Bejikian, Vandhana Krishnan, Lettie McGuire, Xiao Li, Amir Bahmani, and Michael P. Snyder. Real-time alerting system for COVID-19 and other stress events

using wearable data. *Nature Medicine*, 28(1): 175–184, January 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01593-2.

Robert Avram, Geoffrey H. Tison, Kirstin Aschbacher, Peter Kuhar, Eric Vittinghoff, Michael Butzner, Ryan Runge, Nancy Wu, Mark J. Pletcher, Gregory M. Marcus, and Jeffrey Olgin. Real-world heart rate norms in the Health eHeart study. *npj Digital Medicine*, 2(1):1–10, June 2019. ISSN 2398-6352. doi: 10.1038/s41746-019-0134-9.

Gireesh K. Bogu and Michael P. Snyder. Deep learning-based detection of COVID-19 using wearables data, January 2021. ISSN 2124-9474.

Timo B. Brakenhoff, Brianna Mae Goodale, Marcel Van Willigen, Andjela Markovic, Vladimir Kovacevic, Duco Veen, Marianna Mitratza, Janneke van de Wijgert, Billy Franks, Santiago Montes, Eskild K. Fredslund, Serkan Korkmaz, Theo Rispens, Lorenz Risch, Ariel V. Dowling, Amos A. Folarin, Patricia Bruijning, Richard Dobson, Tessa Heikamp, Paul Klaver, Xi Bai, Kirsten Grossman, Weideli Ornella, Paul Klaver, Maureen Cronin, Diederick E. Grobbee, and On behalf of COVID-RED Consortium. Remote Early Detection of SARS-CoV-2 infections (COVID-RED), August 2023.

Tiffany Tianhui Cai, Hongseok Namkoong, and Steve Yadlowsky. Diagnosing Model Performance Under Distribution Shift, July 2023.

Eduardo Casilari and Carlos A. Silva. An analytical comparison of datasets of Real-World and simulated falls intended for the evaluation of wearable fall alerting systems. *Measurement*, 202:111843, October 2022. ISSN 0263-2241. doi: 10.1016/j.measurement.2022.111843.

Sidhartha Chaudhury, Chenggang Yu, Ruifeng Liu, Kamal Kumar, Samantha Hornby, Christopher Duplessis, Joel M. Sklar, Judith E. Epstein, and Jaques Reifman. Wearables Detect Malaria Early in a Controlled Human-Infection Study. *IEEE Transactions on Biomedical Engineering*, 69(6): 2119–2129, June 2022. ISSN 1558-2531. doi: 10.1109/TBME.2021.3137756.

Adam M. Chekroud, Matt Hawrilenko, Hieronimus Loho, Julia Bondar, Ralitza Gueorguieva, Alkomiet Hasan, Joseph Kambeitz, Philip R. Corlett, Nikolaos Koutsouleris, Harlan M. Krumholz, John H. Krystal, and Martin Paulus. Illusory generalizability of clinical prediction models. *Science*, 383(6679):164–167, January 2024. doi: 10.1126/science.adg8538.

Peter Jaeho Cho, Jaehan Yi, Ethan Ho, Md Mobashir Hasan Shandhi, Yen Dinh, Aneesh Patil, Leatrice Martin, Geetika Singh, Brinnae Bent, Geoffrey Ginsburg, Matthew Smuck, Christopher Woods, Ryan Shaw, and Jessilyn Dunn. Demographic Imbalances Resulting From the Bring-Your-Own-Device Study Design. *JMIR mHealth and uHealth*, 10(4):e29510, April 2022. ISSN 2291-5222. doi: 10.2196/29510.

Jennifer L. Cleary, Yu Fang, Srijan Sen, and Zhenke Wu. 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, December 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0277350.

Bryan Conroy, Ikaro Silva, Golbarg Mehraei, Robert Damiano, Brian Gross, Emmanuele Salvati, Ting Feng, Jeffrey Schneider, Niels Olson, Anne G. Rizzo, Catherine M. Curtin, Joseph Frassica, and Daniel C. McFarlane. Real-time infection prediction with wearable physiological monitoring and AI to aid military workforce readiness during COVID-19. *Scientific Reports*, 12(1):3797, March 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-07764-6.

J A Cook and G S Collins. The rise of big clinical databases. *British Journal of Surgery*, 102(2):e93–e101, January 2015. ISSN 0007-1323. doi: 10.1002/bjs.9723.

Shakti Davis, Lauren Milechin, Tejash Patel, Mark Hernandez, Greg Ciccarelli, Siddharth Samsi, Lisa Hensley, Arthur Goff, John Trefry, Sara Johnston, Bret Purcell, Catherine Cabrera, Jack Fleischman, Albert Reuther, Kajal Claypool, Franco Rossi, Anna Honko, William Pratt, and Albert Swiston. Detecting Pathogen Exposure During the Non-symptomatic Incubation Period Using Physiological Data: Proof of Concept in Non-human Primates. *Frontiers in Physiology*, 12:691074, 2021. ISSN 1664-042X. doi: 10.3389/fphys.2021.691074.

Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H. Granat, Tom White, Vincent T. van Hees, Michael I. Trenell, Christoper G. Owen, Stephen J.

Preece, Rob Gillions, Simon Sheard, Tim Peakman, Soren Brage, and Nicholas J. Wareham. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLOS ONE*, 12(2):e0169649, February 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0169649.

Jessilyn Dunn, Mobashir Hasan Shandhi, Peter Cho, Ali Roghanizad, Karnika Singh, Will Wang, Oana Enache, Amanda Stern, Rami Sbahi, Bilge Tatar, Sean Fiscus, Qi Xuan Khoo, Yvonne Kuo, Xiao Lu, Joseph Hsieh, Alena Kalodzitsa, Amir Bahmani, Arash Alavi, Utsab Ray, Michael Snyder, Geoffrey Ginsburg, Dana Pasquale, Christopher Woods, and Ryan Shaw. 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, April 2022. doi: 10.21203/rs.3.rs-1490524/v1.

Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised Visual Domain Adaptation Using Subspace Alignment. In *2013 IEEE International Conference on Computer Vision*, pages 2960–2967, Sydney, Australia, December 2013. IEEE. ISBN 978-1-4799-2840-8. doi: 10.1109/ICCV.2013.368.

Daniel Fuller, Emily Colwell, Jonathan Low, Kassia Orychock, Melissa Ann Tobin, Bo Simango, Richard Buote, Desiree Van Heerden, Hui Luan, Kimberley Cullen, Logan Slade, and Nathan G A Taylor. 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, September 2020. ISSN 2291-5222. doi: 10.2196/18694.

Matteo Gadaleta, Jennifer M. Radin, Katie Baca-Motes, Edward Ramos, Vik Kheterpal, Eric J. Topol, Steven R. Steinhubl, and Giorgio Quer. Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. *npj Digital Medicine*, 4(1):1–10, December 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00533-1.

Hassan M. K. Ghomrawi, Megan K. O'Brien, Michela Carter, Rebecca Macaluso, Rushmin Khazanchi, Michael Fanton, Christopher DeBoer, Samuel C. Linton, Suhail Zeineddin, J. Benjamin Pitt, Megan Bouchard, Angie Figueroa, Soyang Kwon, Jane L. Holl, Arun Jayaraman, and Fizan Abdullah. Applying machine learning to consumer wearable data for the early detection of complications after pediatric appendectomy. *NPJ Digital Medicine*, 6:148, August 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00890-z.

Peter Glynn and Philip Greenland. Contributions of the UK biobank high impact papers in the era of precision medicine. *European Journal of Epidemiology*, 35(1):5–10, January 2020. ISSN 1573-7284. doi: 10.1007/s10654-020-00606-7.

Craig J. Goergen, MacKenzie J. Tweardy, Steven R. Steinhubl, Stephan W. Wegerich, Karnika Singh, Rebecca J. Mieloszyk, and Jessilyn Dunn. Detection and Monitoring of Viral Infections via Wearable Devices and Biometric Data. *Annual Review of Biomedical Engineering*, 24(1):null, 2022. doi: 10.1146/annurev-bioeng-103020-040136.

Jessica R. Golbus, Nicole A. Pescatore, Brahmajee K. Nallamothu, Nirav Shah, and Sachin Kheterpal. 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, November 2021. ISSN 2589-7500. doi: 10.1016/S2589-7500(21)00138-2.

Nir Goldstein, Arik Eisenkraft, Carlos J. Arguello, Ge Justin Yang, Efrat Sand, Arik Ben Ishay, Roei Merin, Meir Fons, Romi Littman, Dean Nachman, and Yftach Gepner. 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, January 2021. ISSN 2077-0383. doi: 10.3390/jcm10215202.

Emilia Grzesiak, Brinnae Bent, Micah T. McClain, Christopher W. Woods, Ephraim L. Tsalik, Bradly P. Nicholson, Timothy Veldman, Thomas W. Burke, Zoe Gardener, Emma Bergstrom, Ronald B. Turner, Christopher Chiu, P. Murali Doraiswamy, Alfred Hero, Ricardo Henao, Geoffrey S. Ginsburg, and Jessilyn Dunn. Assessment of the Feasibility of Using Noninvasive Wearable Biometric Monitoring Sensors to Detect Influenza and the Common Cold Before Symp-

tom Onset. *JAMA Network Open*, 4(9):e2128534, September 2021. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2021.28534.

Gaël P Hammer, Jean-Baptist du Prel, and Maria Blettner. Avoiding Bias in Observational Studies. *Deutsches Ärzteblatt International*, 106(41): 664–668, October 2009. ISSN 1866-0452. doi: 10.3238/arztebl.2009.0664.

Shayan Hassantabar, Novati Stefano, Vishweshwar Ghanakota, Alessandra Ferrari, Gregory N. Nicola, Raffaele Bruno, Ignazio R. Marino, Kenza Hamidouche, and Niraj K. Jha. CovidDeep: SARS-CoV-2/COVID-19 Test Based on Wearable Medical Sensors and Efficient Neural Networks, October 2020.

Robert P. Hirten, Matteo Danieletto, Lewis Tomalin, Katie Hyewon Choi, Micol Zweig, Eddye Golden, Sparshdeep Kaur, Drew Helmus, Anthony Biello, Renata Pyzik, Alexander Charney, Riccardo Miotto, Benjamin S. Glicksberg, Matthew Levin, Ismail Nabeel, Judith Aberg, David Reich, Dennis Charney, Erwin P. Bottinger, Laurie Keefer, Mayte Suarez-Farinas, Girish N. Nadkarni, and Zahi A. Fayad. 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, February 2021. doi: 10.2196/26107.

Robert P Hirten, Lewis Tomalin, Matteo Danieletto, Eddye Golden, Micol Zweig, Sparshdeep Kaur, Drew Helmus, Anthony Biello, Renata Pyzik, Erwin P Bottinger, Laurie Keefer, Dennis Charney, Girish N Nadkarni, Mayte Suarez-Farinas, and Zahi A Fayad. Evaluation of a machine learning approach utilizing wearable data for prediction of SARS-CoV-2 infection in healthcare workers. *JAMIA Open*, 5(2):ooac041, July 2022. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooac041.

Alistair E. W. Johnson, Tom J. Pollard, and Tristan Naumann. Generalizability of predictive models for intensive care unit patients, December 2018.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research, December 2021.

Arinbjörn Kolbeinsson, Piyusha Gade, Raghu Kainkaryam, Filip Jankovic, and Luca Foschini. Self-supervision of wearable sensors time-series data for influenza detection. *arXiv:2112.13755 [cs]*, December 2021.

M. R. Sundara Kumar, D. Salangai Nayagi, Kirubasri G, and Sankar S. A Framework for Detection and Monitoring of COVID-19 using IoT Environment in Pre-Pandemic Life. *International Journal of Computing and Digital Systems*, February 2023. ISSN 2210-142X. doi: 10.12785/ijcds/130159.

B. Naga Lakshmi and M. Robinson Joel. IoT based Illness Prediction System using Machine Learning. In *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1087–1091, March 2023. doi: 10.1109/ICEARS56392.2023.10085553.

Benjamin Lam, Michael Catt, Sophie Cassidy, Jaume Bacardit, Philip Darke, Sam Butterfield, Ossama Alshabrawy, Michael Trenell, and Paolo Missier. 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, March 2021. doi: 10.2196/23364.

Karen Larimer, Stephan Wegerich, Joel Splan, David Chestek, Heather Prendergast, and Terry Vanden Hoek. 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, May 2021. ISSN 1929-0748. doi: 10.2196/27271.

Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex C. Kot. Domain Generalization for Medical Imaging Classification with Linear-Dependency Regularization, October 2020.

Junjie Li, Huaiyu Zhu, Jiaxiang Li, Haotian Wang, Bo Wang, Wei Luo, and Yun Pan. 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, 2023. ISSN 1558-0210. doi: 10.1109/TNSRE.2023.3306203.

Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular

Datasets. *37th Conference on Neural Information Processing Systems*, 2023.

Luca Lonini, Nicholas Shawen, Olivia Botonis, Michael Fanton, Chadrasekaran Jayaraman, Chaithanya Krishna Mummidisetty, Sung Yul Shin, Claire Rushin, Sophia Jenz, Shuai Xu, John A. Rogers, and Arun Jayaraman. 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, 2021. ISSN 2168-2372. doi: 10.1109/JTEHM.2021.3058841.

Carissa A. Low, Anind K. Dey, Denzil Ferreira, Thomas Kamarck, Weijing Sun, Sangwon Bae, and Afsaneh Doryab. Estimation of Symptom Severity During Chemotherapy From Passively Sensed Data: Exploratory Study. *Journal of Medical Internet Research*, 19(12):e9046, December 2017. doi: 10.2196/jmir.9046.

David Madigan, Patrick B. Ryan, Martijn Schuemie, Paul E. Stang, J. Marc Overhage, Abraham G. Hartzema, Marc A. Suchard, William DuMouchel, and Jesse A. Berlin. Evaluating the Impact of Database Heterogeneity on Observational Study Results. *American Journal of Epidemiology*, 178 (4):645–651, August 2013. ISSN 0002-9262. doi: 10.1093/aje/kwt010.

Madhur Mangalam, Arash Sadri, Junichiro Hayano, Eiichi Watanabe, Ken Kiyono, and Damian G. Kelty-Stephen. Reproducible biomarkers: Leveraging nonlinear descriptors in the face of nonergodicity, May 2023.

Hiral Master, Jeffrey Annis, Shi Huang, Joshua A. Beckman, Francis Ratsimbazafy, Kayla Marginean, Robert Carroll, Karthik Natarajan, Frank E. Harrell, Dan M. Roden, Paul Harris, and Evan L. Brittain. 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, November 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-02012-w.

Caleb Mayer, Jonathan Tyler, Yu Fang, Christopher Flora, Elena Frank, Muneesh Tewari, Sung Won Choi, Srijan Sen, and Daniel B. Forger. Consumer-grade wearables identify changes in multiple physiological systems during COVID-19 disease progression. *Cell Reports Medicine*, 3(4):100601, April 2022. ISSN 2666-3791. doi: 10.1016/j.xcrm.2022.100601.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334.

Mike A. Merrill and Tim Althoff. Self-supervised Pretraining and Transfer Learning Enable Flu and COVID-19 Predictions in Small Mobile Sensing Datasets, June 2022.

Mike A. Merrill, Esteban Safranchik, Arinbjörn Kolbeinsson, Piyusha Gade, Ernesto Ramirez, Ludwig Schmidt, Luca Foschini, and Tim Althoff. 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, June 2023.

Aziz Mezlini, Allison Shapiro, Eric J. Daza, Eamon Caddigan, Ernesto Ramirez, Tim Althoff, and Luca Foschini. Estimating the Burden of Influenza-like Illness on Daily Activity at the Population Scale Using Commercial Wearable Sensors. *JAMA Network Open*, 5(5):e2211958, May 2022. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2022.11958.

Dean J. Miller, John V. Capodilupo, Michele Lastella, Charli Sargent, Gregory D. Roach, Victoria H. Lee, and Emily R. Capodilupo. Analyzing changes in respiratory rate to predict the risk of COVID-19 infection. *PLOS ONE*, 15(12):e0243693, December 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0243693.

Tejaswini Mishra, Meng Wang, Ahmed A. Metwally, Gireesh K. Bogu, Andrew W. Brooks, Amir Bahmani, Arash Alavi, Alessandra Celli, Emily Higgs, Orit Dagan-Rosenfeld, Bethany Fay, Susan Kirkpatrick, Ryan Kellogg, Michelle Gibson, Tao Wang, Erika M. Hunting, Petra Mamic, Ariel B. Ganz, Benjamin Rolnik, Xiao Li, and Michael P. Snyder. Pre-symptomatic detection of COVID-19 from

smartwatch data. *Nature Biomedical Engineering*, 4(12):1208–1220, December 2020. ISSN 2157-846X. doi: 10.1038/s41551-020-00640-6.

Marianna Mitratza, Brianna Mae Goodale, Aizhan Shagadatova, Vladimir Kovacevic, Janneke van de Wijgert, Timo B Brakenhoff, Richard Dobson, Billy Franks, Duco Veen, Amos A Folarin, Pieter Stolk, Diederick E Grobbee, Maureen Cronin, and George S Downward. The performance of wearable sensors in the detection of SARS-CoV-2 infection: A systematic review. *The Lancet. Digital Health*, 4(5):e370–e383, May 2022. ISSN 2589-7500. doi: 10.1016/S2589-7500(22)00019-X.

Miho Miyawaki, Walid Brahim, Yosuke Iida, and Jianhua Ma. Recognition of Psychological Stress Levels Using Wearable Biosensors. *International Symposium on Affective Science and Engineering*, ISASE2023:1–4, 2023. doi: 10.5057/isase.2023-C 000027.

Nilah Ravi Nair, Lena Schmid, Fernando Moya Rueda, Markus Pauly, Gernot A. Fink, and Christopher Reining. Dataset Bias in Human Activity Recognition, January 2023.

Aravind Natarajan, Hao-Wei Su, and Conor Heneghan. Assessment of physiological signs associated with COVID-19 measured using wearable devices. *npj Digital Medicine*, 3(1):1–8, November 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0 0363-7.

Larry R. Nelson, Larry A. Nelson, and Leonard D. Zaichkowsky. A Case for Using Multiple Regression Instead of ANOVA in Educational Research. *The Journal of Experimental Education*, 47(4):324–330, 1979. ISSN 0022-0973.

Bret Nestor, Jaryd Hunter, Raghu Kainkaryam, Erik Drysdale, Jeffrey B. Inglis, Allison Shapiro, Sujay Nagaraj, Marzyeh Ghassemi, Luca Foschini, and Anna Goldenberg. Dear Watch, Should I Get a COVID-19 Test? Designing deployable machine learning for wearables, May 2021.

Bret Nestor, Jaryd Hunter, Raghu Kainkaryam, Erik Drysdale, Jeffrey B. Inglis, Allison Shapiro, Sujay Nagaraj, Marzyeh Ghassemi, Luca Foschini, and Anna Goldenberg. Machine learning COVID-19 detection from wearables. *The Lancet Digital Health*, 5(4):e182–e184, April 2023. ISSN 2589-7500. doi: 10.1016/S2589-7500(23)00045-6.

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

Gerald Norman Pho, Nina Thigpen, Shyamal Patel, and Hal Tily. Feasibility of Measuring Physiological Responses to Breakthrough Infections and COVID-19 Vaccine Using a Wearable Ring Sensor. *Digital Biomarkers*, 7(1):1–6, March 2023. ISSN 2504-110X. doi: 10.1159/000528874.

Arvind Pillai, Subigya Kumar Nepal, Weichen Wang, Matthew Nemesure, Michael Heinz, George Price, Damien Lekkas, Amanda C. Collins, Tess Griffin, Benjamin Buck, Sarah Masud Preum, Trevor Cohen, Nicholas C. Jacobson, Dror Ben-Zeev, and Andrew Campbell. 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, December 2023. ISSN 2474-9567. doi: 10.1145/3631452.

Giorgio Quer, Jennifer M. Radin, Matteo Gadaleta, Katie Baca-Motes, Lauren Ariniello, Edward Ramos, Vik Kheterpal, Eric J. Topol, and Steven R. Steinhubl. Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nature Medicine*, 27(1):73–77, January 2021. ISSN 1546-170X. doi: 10.1038/s41591-020-1123-x.

Giorgio Quer, Matteo Gadaleta, Jennifer M. Radin, Kristian G. Andersen, Katie Baca-Motes, Edward Ramos, Eric J. Topol, and Steven R. Steinhubl. Inter-individual variation in objective measure of reactogenicity following COVID-19 vaccination via smartwatches and fitness bands. *npj Digital Medicine*, 5(1):1–9, April 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00591-z.

Jennifer M Radin, Nathan E Wineinger, Eric J Topol, and Steven R Steinhubl. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: A population-based study. *The Lancet Digital Health*, 2(2):e85–e93, February 2020. ISSN 2589-7500. doi: 10.1016/S2589-7500(19)30222-5.

Chaitra Rao, Elena Di Lascio, David Demanse, Nell Marshall, Monika Sopala, and Valeria De Luca. As-

sociation 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, 2023. ISSN 2673-253X.

Dylan M. Richards, MacKenzie J. Tweardy, Steven R. Steinhubl, David W. Chestek, Terry L. Vanden Hoek, Karen A. Larimer, and Stephan W. Wegerich. Wearable sensor derived decompensation index for continuous remote monitoring of COVID-19 diagnosed patients. *npj Digital Medicine*, 4(1):1–11, November 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00527-z.

Martin Risch, Kirsten Grossmann, Stefanie Aeschbacher, Ornella C. Weideli, Marc Kovac, Fiona Pereira, Nadia Wohlwend, Corina Risch, Dorothea Hillmann, Thomas Lung, Harald Renz, Raphael Twerenbold, Martina Rothenbühler, Daniel Leibovitz, Vladimir Kovacevic, Andjela Markovic, Paul Klaver, Timo B. Brakenhoff, Billy Franks, Marianna Mitratza, George S. Downward, Ariel Dowling, Santiago Montes, Diederick E. Grobbee, Maureen Cronin, David Conen, Brianna M. Goodale, and Lorenz Risch. 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, May 2022. ISSN 2044-6055, 2044-6055. doi: 10.1136/bmjopen-2021-058274.

Rebecca Roelofs. Measuring Generalization and Overfitting in Machine Learning. 2019.

Mohammad Rostami and Aram Galstyan. Overcoming Concept Shift in Domain-Aware Settings through Consolidated Internal Distributions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9623–9631, June 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i8.26151.

Tabea Schoeler, Doug Speed, Eleonora Porcu, Nicola Pirastu, Jean-Baptiste Pingault, and Zoltán Kutalik. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour*, 7(7):1216–1227, July 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01579-9.

Allison Shapiro, Nicole Marinsek, Ieuan Clay, Benjamin Bradshaw, Ernesto Ramirez, Jae Min, Andrew Trister, Yuedong Wang, Tim Althoff, and Luca Foschini. Characterizing COVID-19 and Influenza Illnesses in the Real World via Person-Generated Health Data. *Patterns*, 2(1):100188, January 2021. ISSN 2666-3899. doi: 10.1016/j.patter.2020.100188.

Jinjoo Shim, Elgar Fleisch, and Filipe Barata. Circadian Rhythm Analysis Using Wearable-Based Accelerometry as a Digital Biomarker of Aging and Healthspan. Preprint, In Review, December 2023.

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

Harvineet Singh, Vishwali Mhasawade, and Rumi Chunara. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. *PLOS Digital Health*, 1 (4):e0000023, April 2022. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000023.

Justyna Skibińska. *Machine Learning-Aided Monitoring and Prediction of Respiratory and Neurodegenerative Diseases Using Wearables*. Omakustanne/Self-published, 2023. ISBN 978-952-03-3181-8.

Benjamin L. Smarr, Kirstin Aschbacher, Sarah M. Fisher, Anoushka Chowdhary, Stephan Dilchert, Karena Puldon, Adam Rao, Frederick M. Hecht, and Ashley E. Mason. Feasibility of continuous fever monitoring using wearable devices. *Scientific Reports*, 10(1):21640, December 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-78355-6.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, Colorado Springs, CO, USA, June 2011. IEEE. ISBN 978-1-4577-0394-2. doi: 10.1109/CVPR.2011.5995347.

Julie Vaughn, Avital Baral, Mayukha Vadari, and William Boag. 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)*, 2020.

Peter Vorburger and Abraham Bernstein. Entropy-based Concept Shift Detection. In *Sixth International Conference on Data Mining (ICDM'06)*,

pages 1113–1118, December 2006. doi: 10.1109/ICDM.2006.66.

Sholom Wacholder, Nathaniel Rothman, and Neil Caporaso. 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, July 2000. ISSN 0027-8874. doi: 10.1093/jnci/92.14.1151.

Michael Wainberg, Samuel E. Jones, Lindsay Melhuish Beaupre, Sean L. Hill, Daniel Felsky, Manuel A. Rivas, Andrew S. P. Lim, Hanna M. Ollila, and Shreejoy J. Tripathy. 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, October 2021. ISSN 1549-1676. doi: 10.1371/journal.pmed.1003782.

Marc Wiedermann, Robert Bruckmann, and Dirk Brockmann. Corona-Datenspende - Teildatensatz Vitaldaten, August 2023.

Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–34, December 2022a. ISSN 2474-9567. doi: 10.1145/3569485.

Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret E Morris, Eve Riskin, Jennifer Mankoff, and Anind K Dey. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. *36th Conference on Neural Information Processing Systems*, 2022b.

Kan Yamagami, Akihiro Nomura, Mitsuhiro Kometani, Masaya Shimojima, Kenji Sakata, Soichiro Usui, Kenji Furukawa, Masayuki Takamura, Masaki Okajima, Kazuyoshi Watanabe, and Takashi Yoneda. 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, September 2021. doi: 10.2196/30819.

Lei Zhang, Yanjin Zhu, Mingliang Jiang, Yuchen Wu, Kailian Deng, and Qin Ni. Body Temperature Monitoring for Regular COVID-19 Prevention Based on Human Daily Activity Recognition. *Sensors (Basel, Switzerland)*, 21(22):7540, November 2021. ISSN 1424-8220. doi: 10.3390/s21227540.

Appendix 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)).

Appendix 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 TemPredict 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 TemPredict 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.
- 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).

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

Appendix C. TemPredict Dataset

The TemPredict dataset 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 TemPredict 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 Appendix 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.

Appendix 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 TemPredict 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, TemPredict, and *All of Us* datasets.

Appendix 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 is 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.

Appendix 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

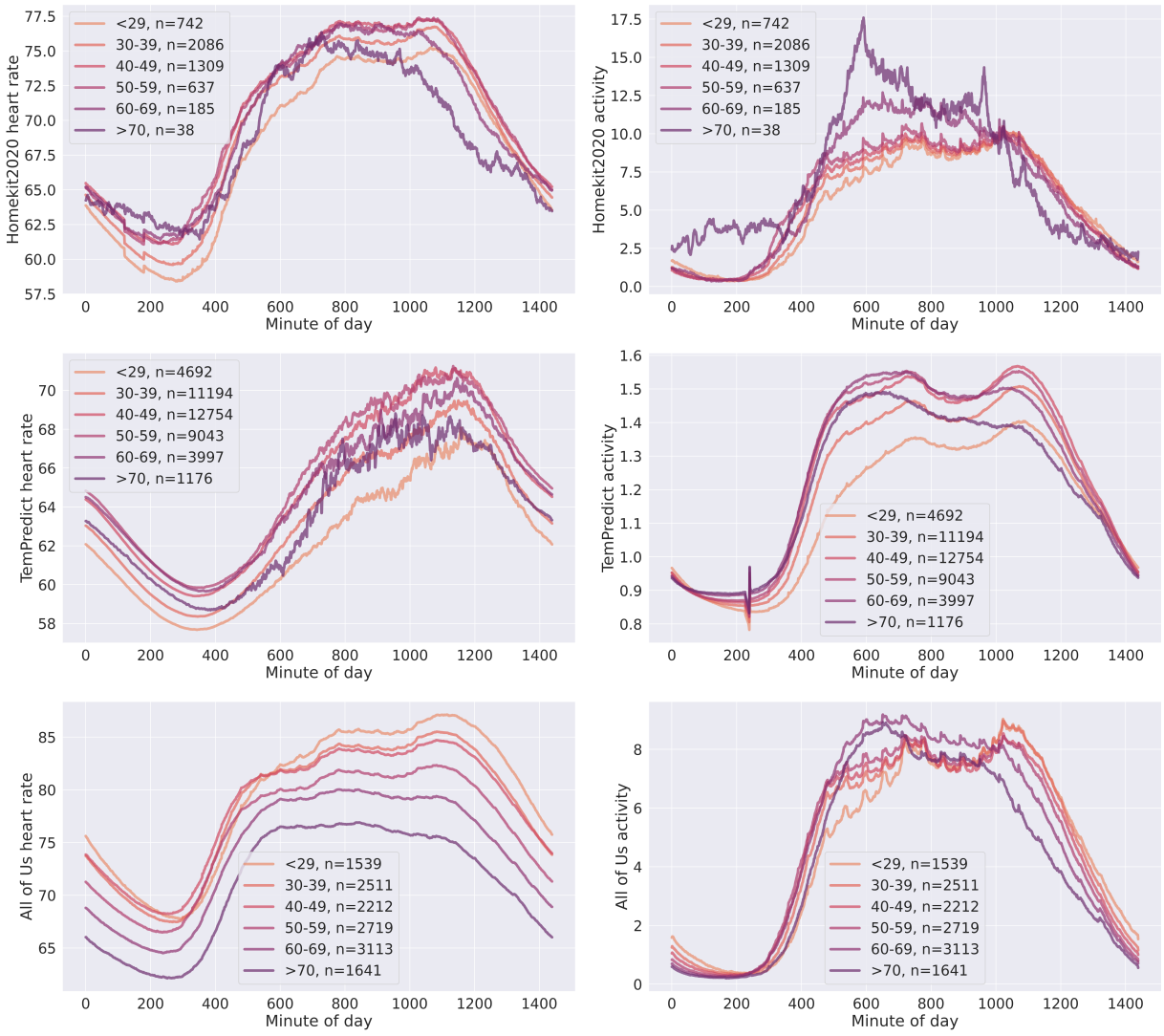

Figure 4: There are within-dataset differences in the mean resting HR based on age throughout the entire day. There also appear to be differences in the patterns of HR and activity throughout the day when comparing across datasets. Lines represent the mean time-of-day wearable-measured average HR (left)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

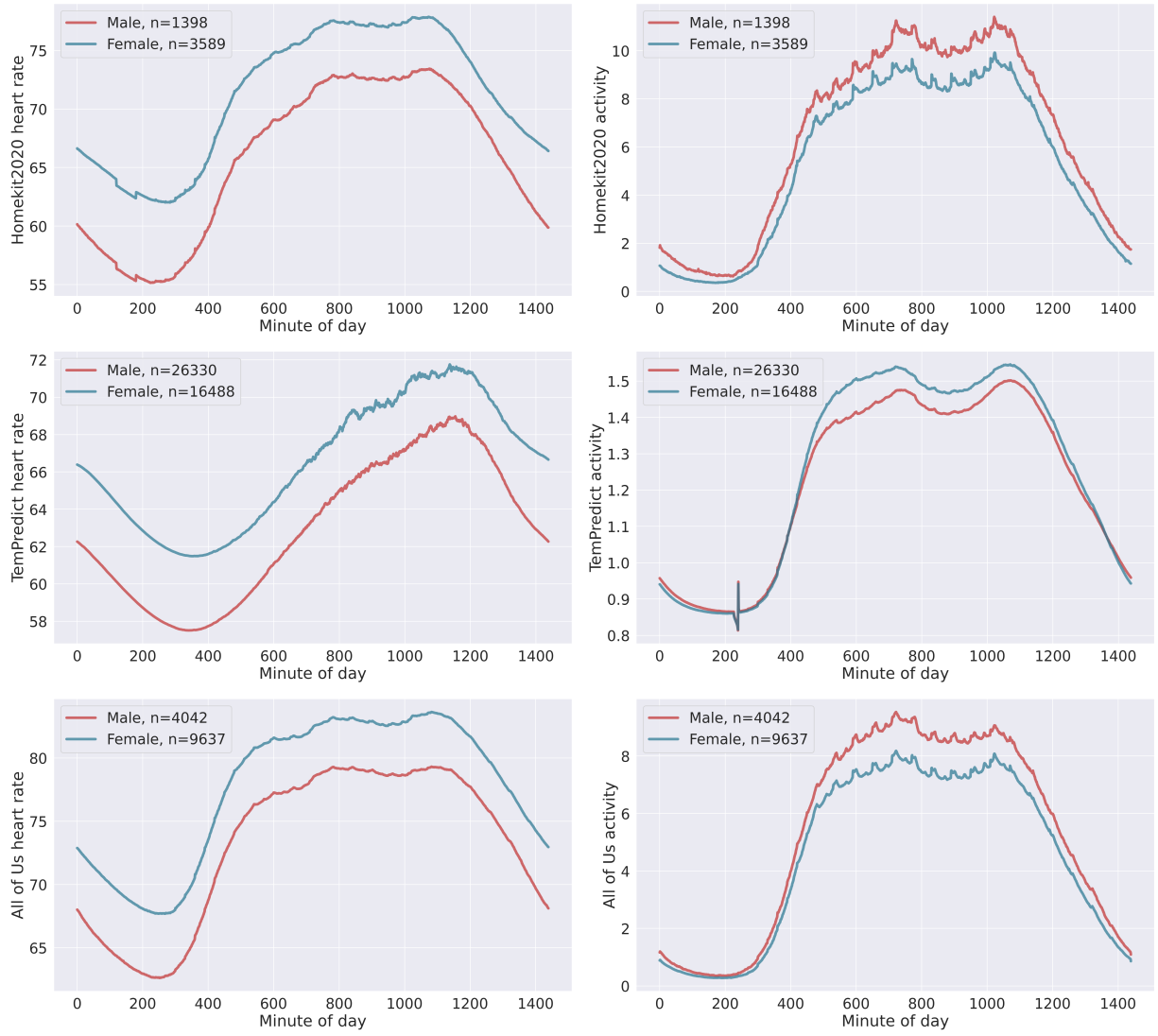

Figure 5: There are within-dataset differences in the mean resting HR based on biological sex throughout the entire day. There also appear to be differences in the patterns of HR and activity throughout the day when comparing across datasets. Time-of-day wearable-measured average heart rate (left) and wearable-measured activity (right). Here, we stratify participants by sex and take the within-dataset mean (top: Homekit, middle: TemPredict, bottom: *All of Us*) for each sex using all available data from that minute of the day.

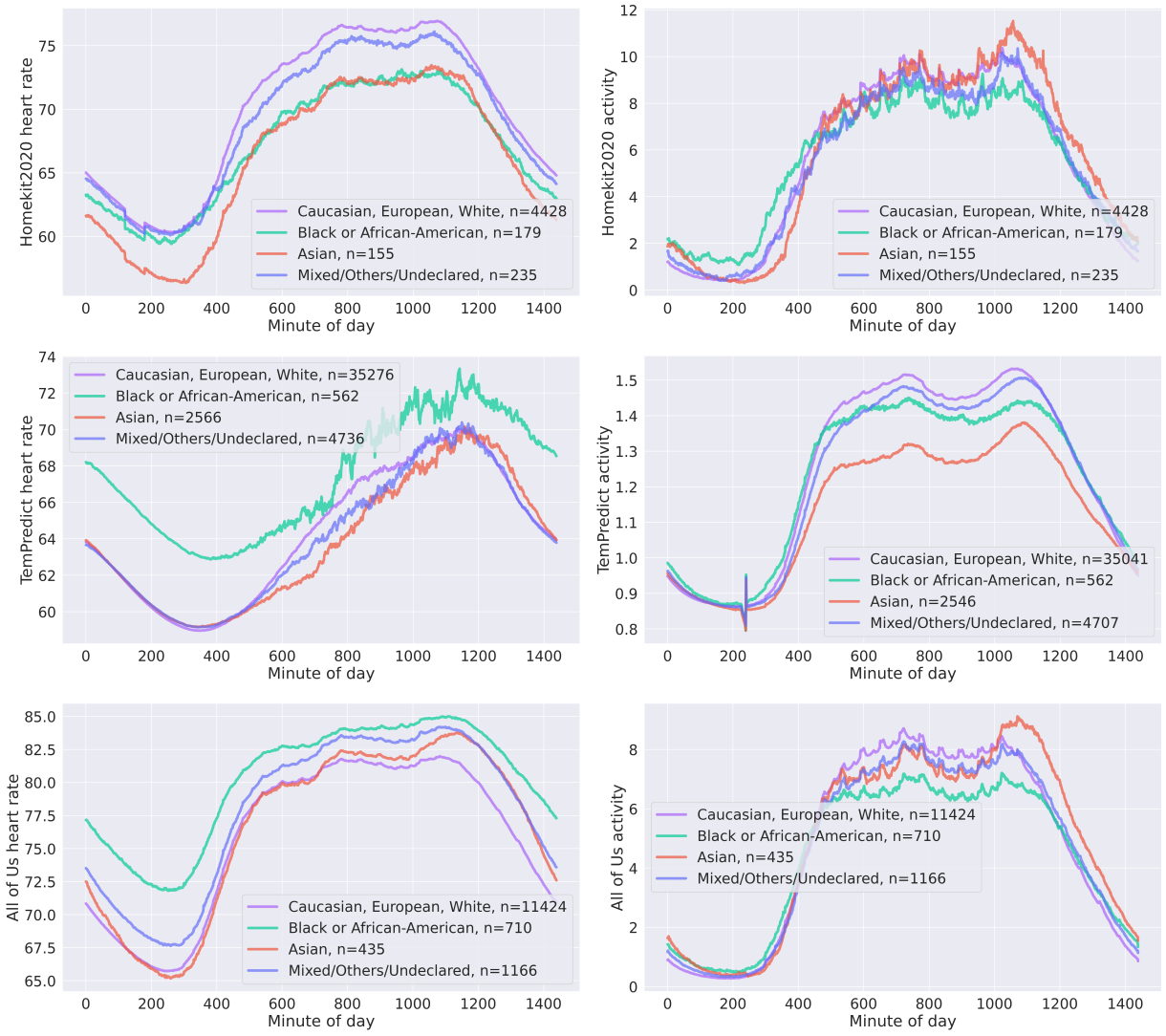

Figure 6: There are within-dataset differences in the mean resting HR based on ethnicity throughout the entire day. There also appear to be differences in the patterns of HR and activity throughout the day when comparing across datasets. Time-of-day wearable measured average heart rate (left) and wearable measured activity (right). Here, we stratify participants by ethnicity and take the within-dataset mean (top: Homekit, middle: TemPredict, bottom: *All of Us*) for each ethnicity using all available data from that minute of the day.

Appendix G. 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 TemPredict (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).

Appendix H. 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.

Appendix I. Participant counts

Table 5: The total number of participants from each dataset whose data are used in Figures 1, 2 and 3 and Tables 2 and 6

	Homekit2020		**TemPredict**		**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

Appendix J. Inter-dataset differences

Table 6: 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 5. Reported as: regression coefficient (p-value).

Age bin	**Sex***	**Dataset†**		
		Homekit	TemPredict	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

Appendix K. Feasibility study review

Table 7: 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 8: 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 9: 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

Appendix L. 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.

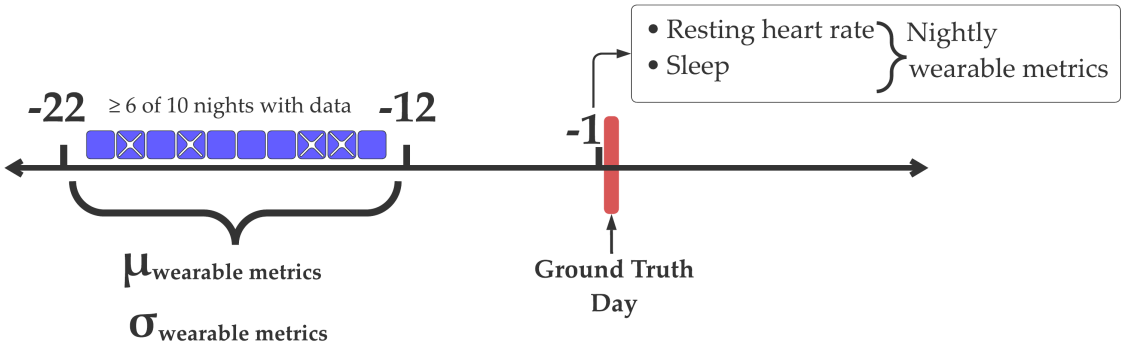

Figure 7: Schematic of the optimized baseline z-score strategy showing an example of how wearable data from the night before the ground truth day is normalized. For detection, data from the night after the ground truth day is z-score normalized by a window that is also shifted forward by one night so that its window is still lagged by twelve days.

Appendix M. 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 7 (ten-day window length with a twelve-day window offset). The prediction model included z-score normalized data from Nights -3, -2, and -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 Appendices B,

C and D. Training and testing follow a stratified five-fold user split cross-validation schema as described in 4.3.6.

Table 10: Performance on prediction tasks across datasets.

Task	Metric \ Dataset	Homekit2020	TemPredict	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 11: Performance on detection tasks across datasets.

Task	Metric \ Dataset	Homekit2020	TemPredict	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 12: Performance of the shared-features model on detection tasks across datasets as measured by average precision (AP).

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

Appendix N. *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*.

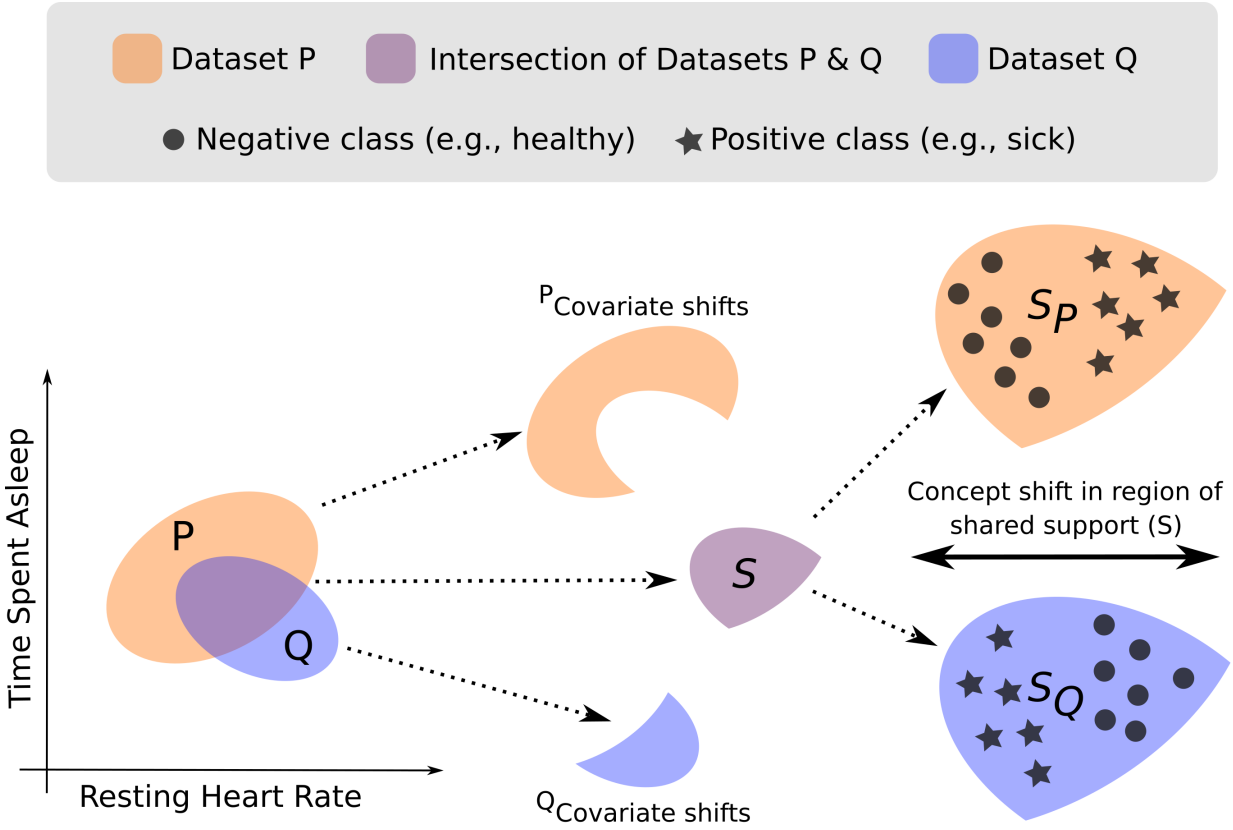

Figure 8: Datasets (distributions) P and Q might both exhibit concept shift and covariate shift. Concept shift can only be estimated in the subsections of feature space which have shared support (e.g., have overlap in their distributions) shown here in purple and labeled as S. The existence of subsections 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.

Appendix O. Normalization aligns dataset means

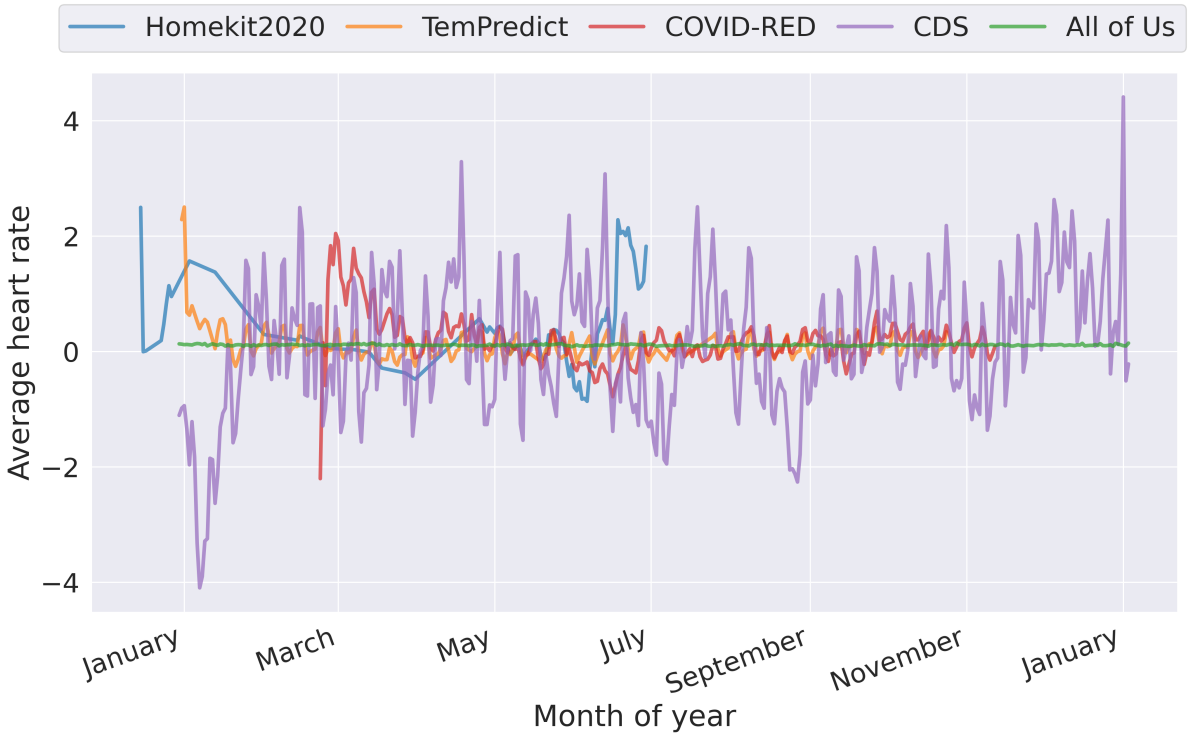

Figure 9: Average daily resting HR taken as the mean across all participants with available z-score normalized data (see Appendix 7) on the same relative date (i.e. 2nd Tuesday of each year) and the mean across repeated relative dates for datasets spanning multiple years (*All of Us* and CDS).