Original Paper

An Algorithm to Classify Real-World Ambulatory Status From a Wearable Device Using Multimodal and Demographically Diverse Data: Validation Study

Sara Popham[1], PhD; Maximilien Burq[1], PhD; Erin E Rainaldi[1], MSc; Sooyoon Shin[1], PhD; Jessilyn Dunn[2,3,4], PhD; Ritu Kapur[1], PhD

[1]Verily Life Sciences, South San Francisco, CA, United States

[2]Department of Biomedical Engineering, Duke University, Durham, NC, United States

[3]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, United States

[4]Duke Clinical Research Institute, Durham, NC, United States

Corresponding Author:
Sara Popham, PhD
Verily Life Sciences
269 E Grand Ave
South San Francisco, CA, 94080
United States
Phone: 1 650 253 0000
Email: spopham@verily.com

Abstract

Background: Measuring the amount of physical activity and its patterns using wearable sensor technology in real-world settings can provide critical insights into health status.

Objective: This study's aim was to develop and evaluate the analytical validity and transdemographic generalizability of an algorithm that classifies binary ambulatory status (yes or no) on the accelerometer signal from wrist-worn biometric monitoring technology.

Methods: Biometric monitoring technology algorithm validation traditionally relies on large numbers of self-reported labels or on periods of high-resolution monitoring with reference devices. We used both methods on data collected from 2 distinct studies for algorithm training and testing, one with precise ground-truth labels from a reference device (n=75) and the second with participant-reported ground-truth labels from a more diverse, larger sample (n=1691); in total, we collected data from 16.7 million 10-second epochs. We trained a neural network on a combined data set and measured performance in multiple held-out testing data sets, overall and in demographically stratified subgroups.

Results: The algorithm was accurate at classifying ambulatory status in 10-second epochs (area under the curve 0.938; 95% CI 0.921-0.958) and on daily aggregate metrics (daily mean absolute percentage error 18%; 95% CI 15%-20%) without significant performance differences across subgroups.

Conclusions: Our algorithm can accurately classify ambulatory status with a wrist-worn device in real-world settings with generalizability across demographic subgroups. The validated algorithm can effectively quantify users' walking activity and help researchers gain insights on users' health status.

(JMIR Biomed Eng 2023;8:e43726) doi: 10.2196/43726

KEYWORDS

digital measurement; wearable sensor; machine learning; ambulatory status; Project Baseline Health Study; physical activity

Introduction

Quantifying physical activity can be highly informative about both general health status and the condition of people with specific diseases [1,2]. Characteristics of physical activity have been shown to be prognostic factors in various chronic conditions [3-13]. Yet reliably producing research-grade measurements of physical activity in real-world settings remains a challenge. Traditionally, the validation of such measurements often relies on individual self-reports or is performed

XSL•FO
RenderX

episodically and in artificial laboratory environments. These approaches suffer from known challenges, such as subjectivity, assessment bias, and unreliability [14-16].

Recently, the advent of wearable technology has made it possible to measure physical activity to a previously untenable extent [17,18]. Ambulatory activity in particular, namely whether individuals are walking and how much, is a basic aspect of physical activity that can be investigated in general populations and in specific clinical settings. Wearable devices can collect information passively during daily living and generate a vast quantity of digital measurements that allow researchers to probe functional physical activity generally and ambulatory activity specifically. Using these digital measures in research studies, however, requires analytical validation [19]. In their design, validation studies have to balance factors such as feasibility and the resource-intensiveness of their data collection approach with demonstrating validity in representative populations.

To date, the majority of measurements in validation studies have come from either short observation periods in laboratory settings [20,21] or self-reported labels in real-world settings [22]. Laboratory measurements often render observations with exceptionally clean and easy-to-use ground-truth labels, but algorithms trained on data of this kind do not always generalize to everyday activities [23]. On the other hand, using self-reported labels as the ground truth yields a closer reflection of individual everyday activities, but these labels are often noisy and less accurate [15,16,24]. There have been some examples of reference devices deployed to generate accurate truth labels in generalizable real-world settings [25,26], but this came at the cost of intrusiveness and resource-intensive data processing steps after collection, such as manual video footage tagging. With all these considerations in mind, validation studies tend to be highly heterogeneous, and need to be interpreted in context.

Herein we report on the development and analytical validation of an ambulatory status classification algorithm. This algorithm classifies the ambulatory status of users of a wrist-worn device in real-world environments. We carried out 2 separate studies including participants from independent populations with distinct sources of ground-truth labels for a deeper characterization of the algorithm performance. One of the studies, the pilot program study, used a relatively small and demographically homogeneous cohort, where participants provided a highly accurate ground-truth source from a reference device. The other study was derived from the Project Baseline Health Study (PBHS), a prospective, multicenter, longitudinal study with participants of diverse backgrounds who were representative of the entire health spectrum [27]; this was a demographically diverse cohort that provided self-reported labels as the ground-truth source. This cohort was also relatively large, and we therefore expected it to yield results less susceptible to outlier readouts. We present analytical validation results of the performance of our algorithm against the highly accurate ground-truth source (from the pilot), and we examine the generalizability of the results across a study population of demographically diverse individuals (in the PBHS).

Methods

Participant Cohorts

Two distinct studies were conducted, with training and testing groups identified a priori within each study. Participants in both studies wore the smartwatch (the Verily Study Watch) [27-30].

The first study was a pilot program (n=75) of adult volunteer participants recruited among Verily Life Sciences employees in 2 locations (South San Francisco, California, and Cambridge, Massachusetts) without specific selection criteria. For this group, ground-truth labels were collected from an ankle-worn reference device (StepWatch 4). The Verily Study Watch and reference device were worn simultaneously for 7 consecutive days to ensure capture of both weekday and weekend behavior; for each participant, days were included as evaluable if both devices were worn synchronously for a minimum of 8 hours. No demographic information on race or ethnicity was collected in this study. The observation period ran from June to December 2019.

In order to expand the demographic representativeness of the overall validation effort, the second study included a large and diverse cohort (n=1691) consisting of participants from the PBHS who consented to participate in this substudy [27]. The period for data collection ran from May to December 2019.

Ethics Approval

The pilot program was determined to be exempt research that did not require institutional review board review. Written informed consent was obtained from all participants enrolled in the PBHS; the PBHS was approved by both the WCG institutional review board (approval tracking number 20170163, work order number 1-1506365-1) and the institutional review boards of each participating institution (Stanford University, Duke University, and the California Health and Longevity Institute) [27]. The PBHS was registered at ClinicalTrials.gov (NCT03154346).

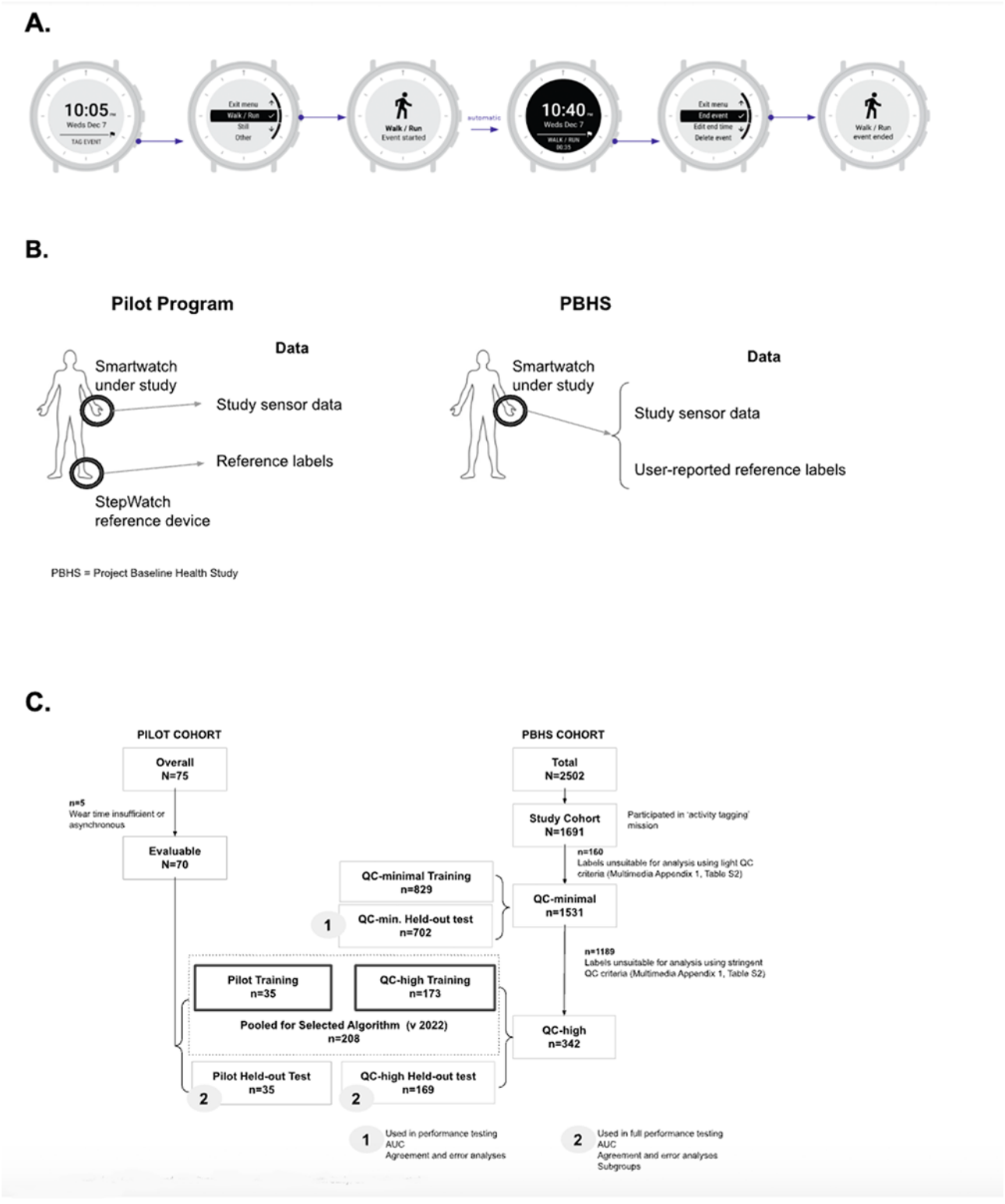
All methods complied with relevant guidelines and regulations; the research involving human participants was performed in accordance with relevant guidelines and regulations. Experimental protocols were approved by appropriate committees from Verily Life Sciences and by PBHS governance (participating institutions are above).

Wearable Devices

The Verily Study Watch recorded acceleration data in both cohorts via an onboard inertial measurement unit with a 30 Hz 3-axis accelerometer. For the PBHS population, the smartwatch also contained a user interface allowing participants to tag their activities on the watch (Figure 1A).

XSL•FO
RenderX

Figure 1. (A) Sketch of the user interface of the study device used in the Project Baseline Health Study. (B) Data elements for the 2 studies. (C) Flow of participant inclusion for the different cohorts and data sets in the 2 studies. AUC: area under the receiver operating characteristic curve; PBHS: Project Baseline Health Study; QC: quality control.



The reference device for the pilot program was an ankle-worn single-axis accelerometer (Modus StepWatch 4) that provided step count as a reference label for algorithm development.

Reference Labels

In the pilot program, we generated reference labels on data collected from the ankle-worn StepWatch: 10-second windows were considered "ambulatory" if they had ≥3 steps on the wearing foot and "nonambulatory" if they had <3 steps [31]. The default window size returned by this device was 10 seconds, and this was deemed to provide good temporal granularity.

For the self-reported reference labels from the PBHS, participants tagged their activities as 1 of 3 options listed by the wrist device: "walk/run," "still," and "other." Participants could tag the start and end of an activity period directly on the

XSL•FO
RenderX

watch, which enabled precise synchronization of the labels to the raw sensor data stream. When necessary, participants could edit or delete tags as needed (Figure 1B). For the purpose of this analysis, "still" and "other" were grouped together under the "nonambulatory" label, while "walk/run" was equated to "ambulatory."

The amount of data used from each of these studies is summarized in Multimedia Appendix 1, Table S3.

Algorithm Development

Data from each study (pilot program, n=75; PBHS, n=1691) were split into nonoverlapping training and testing data sets at the participant level. For each study, data from approximately half the participants were used for training the algorithm and data from the other half were held out for algorithm testing. We decided on a 50-50 split in order to retain statistical power in the testing data, particularly considering the intended additional analyses of different demographic subgroups (discussed below).

In the pilot program, the split into training and testing data sets was based on participants' daily step counts in order to mitigate potential algorithmic biases caused by training primarily on data from participants with either very low or very high activity levels. The difference in the mean daily step counts between the 2 halves of the split was 234 steps. For the PBHS cohort, the split into training and testing data sets was done randomly, as participants did not have daily aggregated results. We trained multiple versions of the algorithm with combinations of different subsets of training data and compared performance across these different algorithms (Figure 1).

We developed an algorithm that classifies the ambulatory status of device users in 10-second epochs (as ambulatory vs nonambulatory). First, the following 14 features were extracted from the Verily Study Watch's acceleration data, in 10-second epochs: 3 features related to deviations of the signal, 5 features derived from power spectral density energy in frequency bands typically associated with user ambulation (ie, walking or running), 2 features that are signal percentiles (ie, 95th percentiles), and 4 features that are differences between signal percentiles (ie, IQR). These features were fed into a shallow neural network model with 2 dense layers: ReLu nonlinearities and softmax of outputs. The neural network was trained with a batch size of 32. The Adam optimizer was used with a learning rate of 0.001, and loss was calculated using categorical cross-entropy. Training ran for 10 epochs. Alternative features and neural network architectures were explored using the training data, but larger feature sets or more complex architectures did not result in higher performance, so this algorithm was chosen.

The classifier threshold was optimized to minimize absolute percentage error on daily ambulatory time on the training data from the pilot study (vs the data from the reference device used as the ground-truth source, as discussed above). For this optimization process, we performed 5-fold cross-validation at the participant level within the training data. We found the minimum daily mean absolute percentage error (MAPE) across the aggregated held-out data from all folds using a 1D grid search procedure.

The signal-processing, feature selection, model training, and hyperparameter tuning were all performed on training data sets identified a priori.

Analyses

The demographic characteristics of the study cohorts were analyzed using descriptive statistics.

We analyzed the following metrics to characterize the performance of the algorithm, calculated on the held-out test sets: area under the receiver operating characteristic curve (AUC) for the overall study cohorts and across different demographic subcohorts within the PBHS cohort (this was chosen as the metric for comparison because, unlike other measures, such as F_1-score or accuracy, it is not susceptible to differences in the chosen classifier threshold), mean accuracy, and MAPE of daily ambulatory time, defined as the summing of all 10-second windows that were labeled as "ambulatory" in a day.

Analyses were performed in python using *NumPy* (version 1.21.5), *pandas* (version 1.1.5), *SciPy* (version 1.2.1), *scikit-learn* (version 1.0.2), and *tensorflow* (version 2.10.0).

Confidence intervals were calculated using the bootstrap method with 1000 resampling iterations. Resampling was done at the participant level to ensure that all data from a single participant were either included or excluded within each resampling iteration.

Results

Characteristics of Participants From the Pilot Study and the PBHS Cohort

Participants in the pilot study were mostly male (45/75, 64%), with a mean age of 33 (SD 8.5) years. Participants from the PBHS were more often female (1366/2502, 55%), with a mean age of 54 (SD 17) years (Multimedia Appendix 1, Table S1).

Algorithm Training

Data from each study were separately split (approximately 50-50) into nonoverlapping training and testing data sets (Figure 1); this allocation was done at the participant level (n=75 from the pilot study and n=1691 from the PBHS population). Out of 16.769 million 10-second epochs collected from the 2 studies, 8.841 million 10-second epochs were used for training across all algorithm iterations generated (the data sets are described in Multimedia Appendix 1, Table S3).

From the pilot program study, a total of 1,641,272 nonoverlapping 10-second epochs were collected (n=70 participants; Figure 1), of which 228,721 (13.9%) were "ambulatory" according to the reference device–based labels. We used 879,593 10-second epochs (from 35 unique participants) for training (118,730, 13.5% of which were "ambulatory"; Multimedia Appendix 1, Table S3).

We collected a total of 14,814,910 nonoverlapping 10-second epochs from the PBHS (n=1531 participants; Figure 1), of which 7,079,216 (47.8%) were "ambulatory" according to the participant-reported reference labels. The proportion of

XSL•FO
RenderX

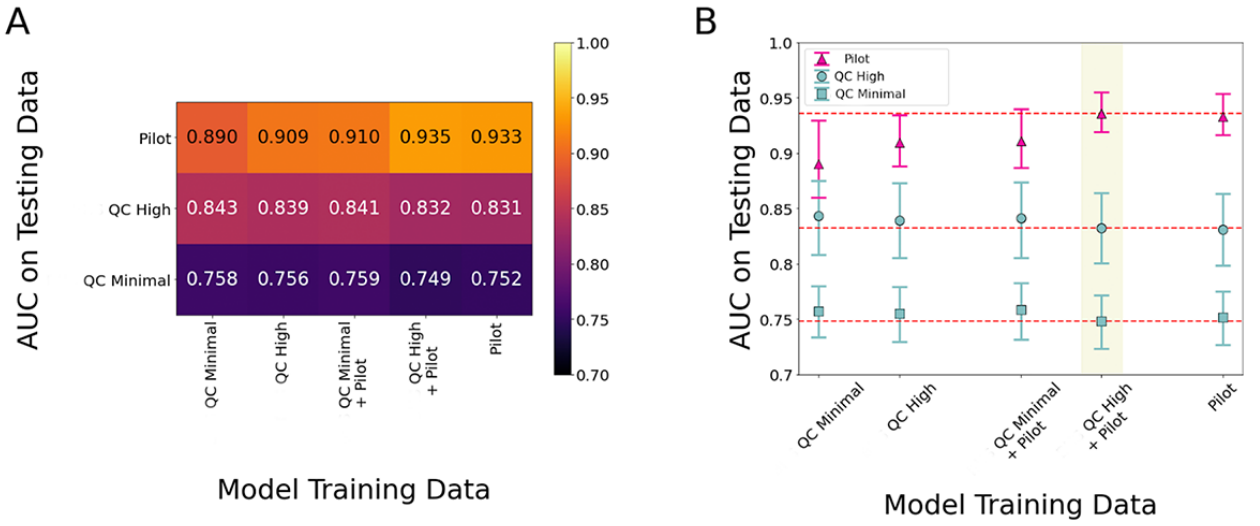
"ambulatory" labels in the PBHS was higher than in the pilot program study (47.8% vs 13.9%), which is likely attributable to the different labeling methods across studies. We expect that labeling from the pilot study was more stringent to show true ambulatory epochs, because these were determined directly by the reference device readouts (ie, any 10-second epoch with greater than or equal to 6 steps, relative to all 10-second epochs collected during the wear time). In the PBHS, the proportion of ambulatory labels was determined based on participant self-reported, manually entered walk/run tags relative to all entered tags. PBHS tagging, therefore, can be more vulnerable to selection bias toward "ambulatory," since participants may favor reporting active over inactive states.

Data from the PBHS were not only divided into training and testing sets, but, across each set, we considered 2 quality control (QC) strata to test the impact of data quality on the development and performance of the algorithm. An extremely light QC selection, eliminating labels with gross apparent user errors (such as tags that were longer than a full day), was applied to generate the "QC-minimal" sub–data set, which therefore included virtually all labels suitable for evaluation (10,264/104,212, 9.8% of user-tagged events were eliminated, and another 12,010/104,212, 11.5% were truncated); a more stringent selection was applied to generate the "QC-high" sub–data set (80,852/104,212, 77.6% of user-tagged events were eliminated, and all tags were truncated to some degree; Figure 1 and Multimedia Appendix 1, Table S2). The 2 strata aimed to parse out performance variability due to noise generated by imperfectly self-reported reference labels (this was not a factor for the labels from the reference device in the pilot program). The resulting size of these QC training sub–data sets was 160,778 10-second epochs for QC-high (n=173 participants) and 7,802,829 for QC-minimal (n=829 participants). Of these labeled epochs, 102,783 (63.7%) and 3,863,964 (49.5%), respectively, were ambulatory according to the participant-reported tags (Figure 1 and Multimedia Appendix 1, Table S3).

Effect of Raw Data Quality on Algorithm Performance

We tested each of the algorithm iterations from the training process above (originated using the 2 PBHS QC sub–data sets and the pilot data set) across data from the held-out QC sub–data sets from the PBHS and the pilot program by calculating AUC values across all combinations. Namely, we tested each of the following algorithms against the held-out data sets from the pilot study and the PBHS QC-high and QC-minimal sub–data sets (Figure 2): (1) trained with the PBHS QC-high sub–data set, (2) trained with the PBHS QC-minimal sub–data set, (3) trained with the pooled PBHS QC-high plus pilot data set, (4) trained with the pooled PBHS QC-minimal plus pilot data set, and (5) trained with just the pilot data set. For each algorithm iteration, AUC values varied across the testing sub–data sets (QC sub–data sets from the PBHS and pilot program), with differences ranging from 0.047 to 0.187. For each test data set, the AUC variations across the algorithm iterations (1) through (5) were narrower, with differences ranging between 0.001 and 0.045. Therefore, data quality differences across the training sub–data sets did not appear to affect algorithm performance, as reflected in AUC variability, as much as data quality in the testing sub–data sets.

Figure 2. (A) Heat map of AUC values for the algorithm iterations generated via different training sub–data sets from the PBHS when tested on each of the separate testing cohorts. (B) AUC values for the algorithm iterations generated via different training sub–data sets from the PBHS when tested on each of the separate testing cohorts, with error bars based on the 95% CI. Each testing cohort is shown with a different color or symbol. From top to bottom, the red dotted lines indicate mean AUC values for the pilot, PBHS QC-high, and PBHS QC-minimal test data sets, respectively. The model trained on combined PBHS QC-high and pilot training data (highlighted in yellow) was the version of the algorithm used for further analyses. AUC: area under the receiver operating characteristic curve; PBHS: Project Baseline Health Study; QC: quality control.



Based on the testing results described above, we selected an algorithm trained using combined data from one of the PBHS sub–data sets (QC-high) plus the pilot program data set to proceed to further analysis. This algorithm iteration (termed "version 2022") showed the highest testing performance (evaluated by AUC) calculated with data from the pilot program (the most precise and cleanest data set) without substantially reduced performance on PBHS data (Figure 2). With this approach, we prioritized testing the accuracy of the algorithm against participants' actual ambulatory status based on the

XSL•FO
RenderX

reference device, not against the type of labels that are most feasible to obtain (ie, self-reported labels), although we report accuracy on both types of labels.

Algorithm Testing

Tested against the held-out data set from the pilot program (Table 1), the selected algorithm had a sensitivity of 71% and a specificity of 95%, for an overall accuracy of 91.5% (95% CI 90.3%-92.9%; Figure 3A) and an AUC of 0.938 (95% CI 0.921-0.958; Figure 3B) when classifying the ambulatory status of 10-second epochs. When tested on the held-out data set from the PBHS QC-high sub–data set, the selected algorithm had an overall accuracy of 75.7% (95% CI 72.5%-78.6%) and an AUC of 0.832 (95% CI 0.800-0.864).

Table 1. Algorithm performance measures.

	Accuracy	Sensitivity	Specificity	PPV[a]	F_1-score	AUC-ROC[b]	AUC-PRC[c]
Pilot study	91.3%	0.706	0.948	0.696	0.701	0.938	0.781
PBHS[d] QC[e]-high	75.8%	0.731	0.802	0.885	0.788	0.832	0.901

[a]PPV: positive predictive value.

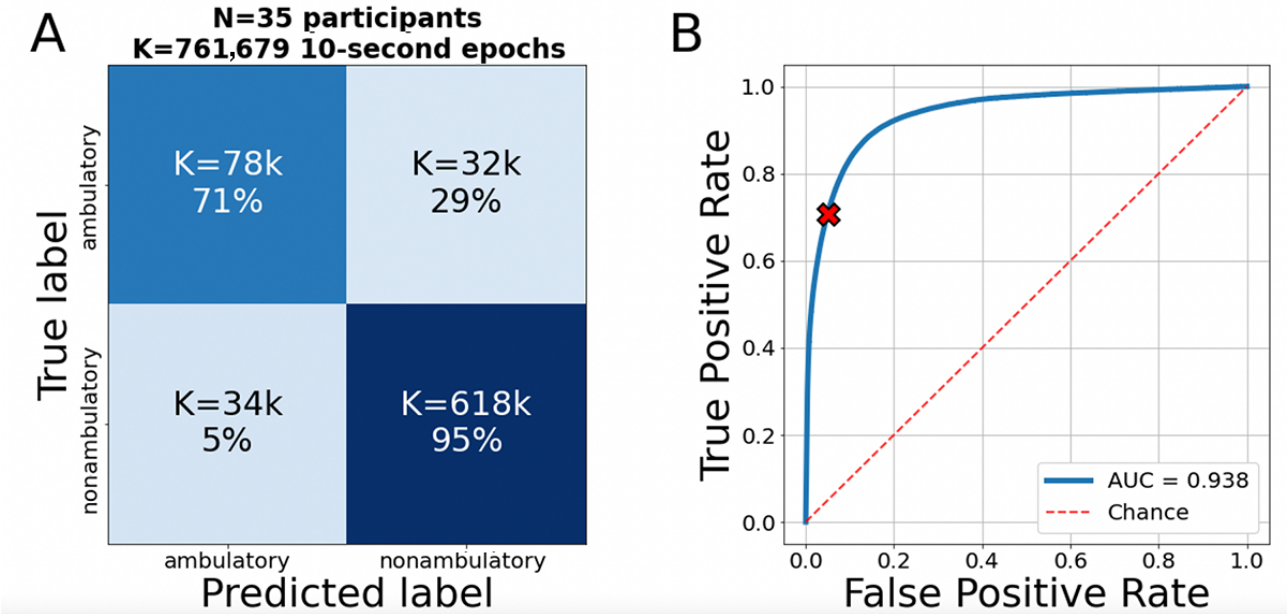
[b]AUC-ROC: area under the receiver operating characteristic curve.

[c]AUC-PRC: area under the precision-recall curve.

[d]PBHS: Project Baseline Health Study.

[e]QC: quality control.

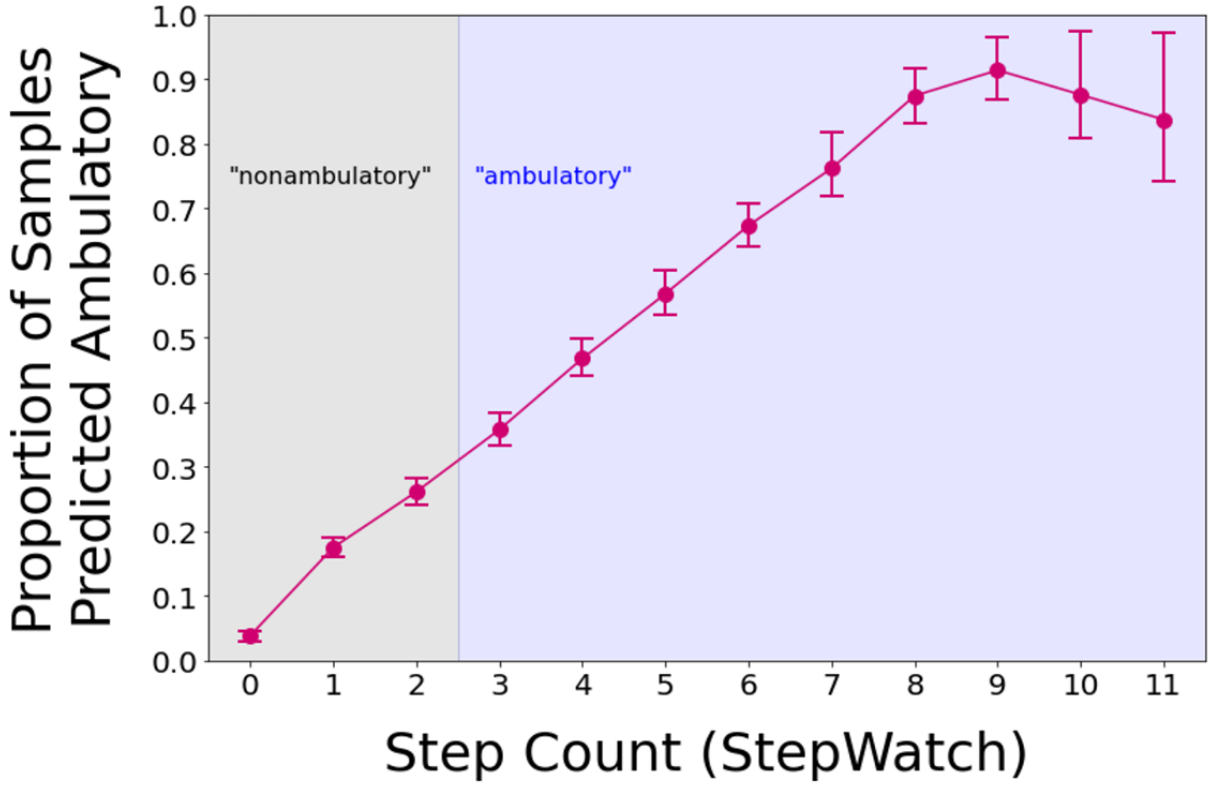
Figure 3. (A) Accuracy of the algorithm selected for full analysis, as evaluated in the pilot cohort. Here, the color map denotes K, the number of 10-second epochs. Percentages are normalized across rows, which allows easy reading of the sensitivity and specificity values. (B) Receiver operating characteristic curve and area under the curve of the algorithm selected for full validation, as evaluated in the pilot cohort. The red X denotes the true positive rate and false positive rate of the algorithm at the chosen classifier threshold. AUC: area under the receiver operating characteristic curve.



The proportion of predicted ambulatory epochs of the selected algorithm varied with the number of steps in the 10-second epochs (Figure 4). The lowest proportion of predicted ambulatory epochs happened in the 3 to 5 step range (36%-57% sensitivity, ie, correct predictions as "ambulatory: yes"), and the proportion of epochs classified as ambulatory grew with additional steps in the 10-second window (67%-91% correct predictions). Note that data from epochs with more than 11 recorded single-leg steps are not shown due to their low frequency (the number of samples per step count is shown in Multimedia Appendix 1, Figure S1).

XSL•FO
RenderX

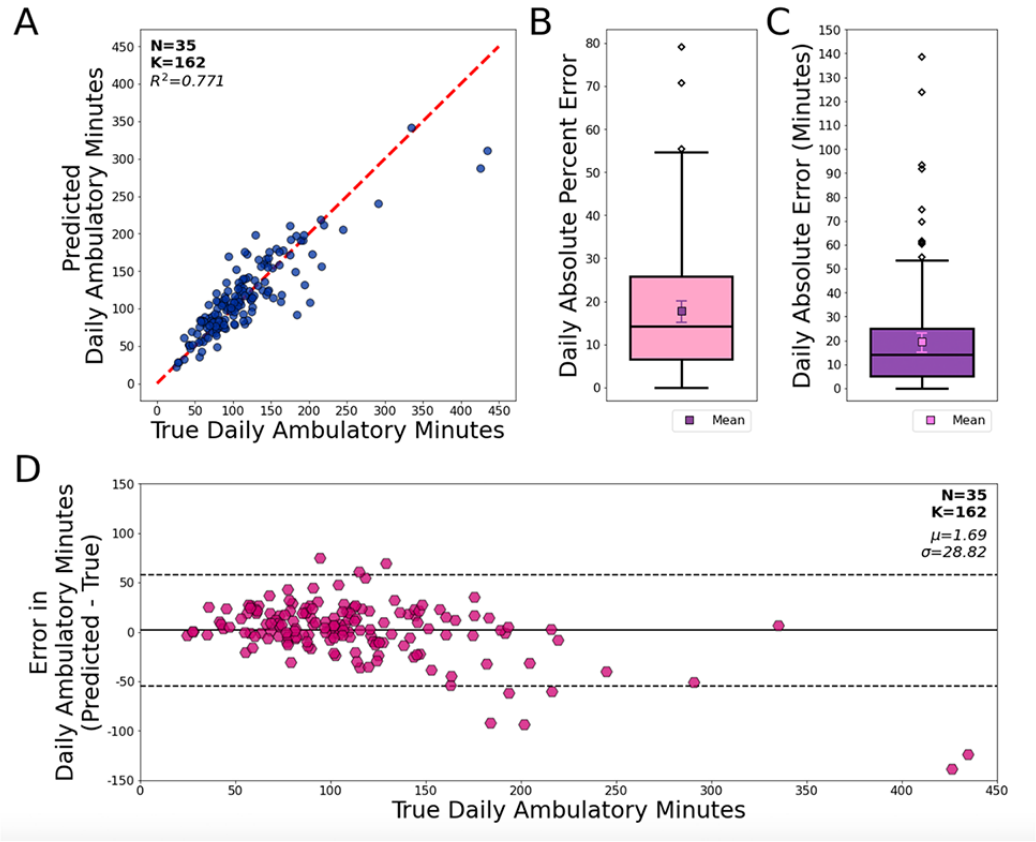
Figure 4. Predictions of the selected algorithm to classify 10-second epochs as ambulatory (or not) according to the number of steps in the 10-second epochs based on the reference device data from the pilot program study. A perfectly performing algorithm would predict "ambulatory" for all epochs with 3 or more steps on the wearing foot (indicated by the blue shadow), and nonambulatory for all epochs with fewer steps (indicated by the gray shadow). Epochs with more than 11 recorded steps are not shown due to their small sample size (Multimedia Appendix 1, Figure S1).



When considering daily step aggregates as the metric of interest, there was good agreement between the algorithm classifications and the reference (R^2=0.771), with a MAPE in daily ambulatory time (minutes) of 18% (95% CI 15%-20%) and a median absolute percentage error of 14% (Figure 5A and Figure 5B). The mean absolute error (MAE) of daily ambulatory time was 19.5 (95% CI 15.0-23.2) minutes, and the median absolute error was 14 minutes (Figure 5C). Consistent with the observations at the 10-second epoch level, the magnitude of error in daily ambulatory time (ie, the difference between algorithm-predicted and actual values) was dependent on the actual daily ambulatory time (as computed by the StepWatch; Figure 5D): the chance for underestimating daily ambulatory time (in minutes) grew as the reference daily ambulatory time increased. The largest underestimation we observed was 138.5 minutes in absolute time (relative error 32.5%).

XSL•FO
RenderX

Figure 5. Agreement and error rates of the algorithm predictions. K is the number of user-days. (A) Agreement between the selected algorithm's predictions and the ground-truth source for daily ambulatory time in the pilot-testing data set. (B) Absolute percent error in daily ambulatory time: median (pink box) and mean (purple box). (C) Absolute error in daily ambulatory time in minutes: median (purple box) and mean (pink box). (D) Modified Bland-Altman plot showing error in daily ambulatory time (in minutes) as it relates to the ground-truth daily ambulatory time.

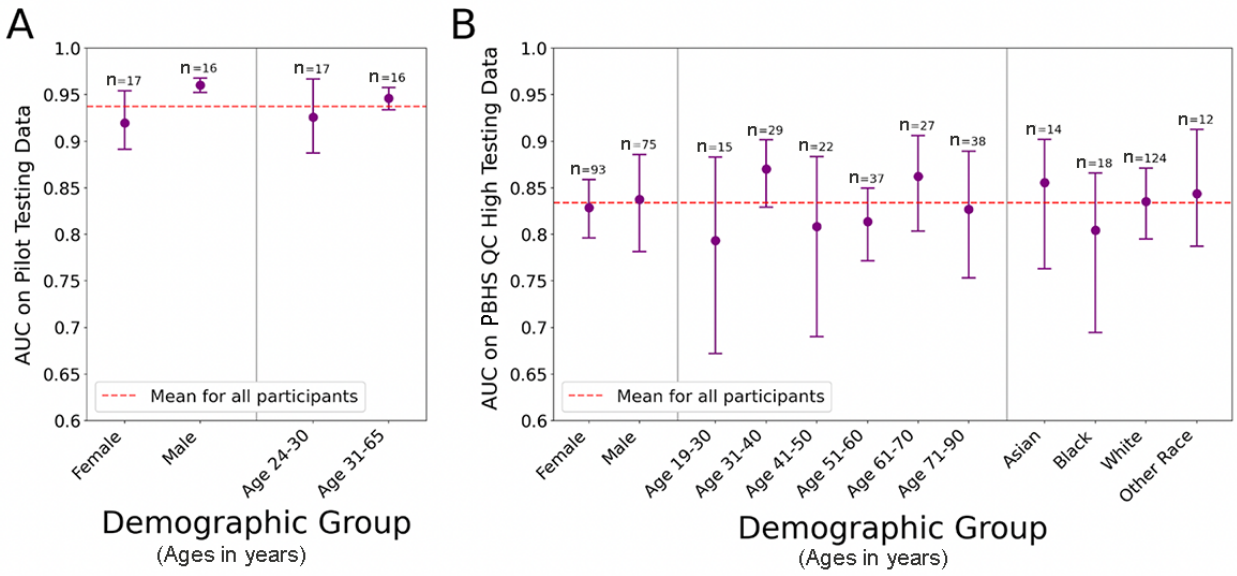


Performance of the Ambulatory Status Classification Algorithm Across Demographic Subgroups

In order to characterize the generalizability of the algorithm's performance, we calculated AUC values for the selected algorithm across demographic subgroups based on gender, age, and race. Initially, in the testing data set from the pilot program (Figure 6A), the results suggested a possible difference in performance between male and female participants, as seen in the lack of overlap of the 95% CIs. However, in a similar analysis using the larger and more diverse testing data set from the PBHS, which enabled subanalyses by participant gender, race, and age, that difference was no longer present and the results showed no meaningful performance difference across any of the subgroups of age, gender, or race, as evidenced by the overlapping 95% CIs (Figure 6B). A replication of the majority population from the pilot study within the PBHS showed an AUC of 0.8166 (95% CI 0.7501-0.8666) for White males aged 31 to 65 years in the PBHS cohort, which was not significantly different from the AUC of the PBHS cohort as a whole (AUC 0.8339).

XSL•FO
RenderX

Figure 6. Performance (AUC values) of the selected algorithm across different demographic subgroups. (A) The pilot study testing data set. (B) The PBHS QC-high testing data set. AUC: area under the receiver operating characteristic curve; PBHS: Project Baseline Health Study; QC: quality control.



Discussion

This study presents the analytical validation in a real-world setting of an algorithm to classify the ambulatory status of users wearing a smartwatch. The algorithm performs well, distinguishing between ambulatory and nonambulatory states with high accuracy (75.7%-91.5% depending on the testing data set). Furthermore, the approach taken to analytic validation allowed us to investigate multiple subgroups, including age, gender, and race, demonstrating that the high performance of the algorithm is generalizable across a broad range of demographics.

All existing validation studies of ambulatory status classification from wrist-worn sensors have been either performed on young and healthy populations [25] or in the laboratory or clinic [20,21]. Yet measuring ambulatory status or daily ambulatory time is most clinically relevant for people with walking impairments—whether due to age, movement disorders, cardiovascular illness, or other circumstances—and most informative when done in an individual's own environment (ie, their real-world setting). Thus, a key innovation in this work is our focus on using data captured in real-world settings (as opposed to highly controlled clinic or laboratory settings) from demographically diverse cohorts for the actual development and validation of this algorithm.

Therefore, the novel contributions of this work are 2-fold. First, we introduce a scalable framework for collecting reference labels on ambulatory status via a reference device and via user-reported data for training and validation. As part of that approach, we used 2 separate and different modalities to measure ground-truth status. This strategy enabled us to handle both comprehensive and highly precise labels (in the pilot program), as well as a larger volume of inherently noisy ones (user-reported tags from the PBHS), both in real-world settings. Our strong results across both sets of data indicate that this innovative multimodal approach contributed to a robust development scheme that may have boosted the performance of the resulting algorithm. The long-term practical convenience of a wrist-worn device (as opposed to an ankle-worn device or a dedicated assessment period) may be advantageous for this type of continuous generalizable monitoring [32-35], although a thorough side-by-side analysis of these 2 reference standard measurement methods to fully understand their correlation remains as a topic for future studies.

Second, we leveraged this framework to provide large-scale validation of the performance of the selected algorithm iteration, addressing shortcomings in terms of generalization previously reported in the literature [20,21,32]. Prior studies have used algorithms to report on differences in physical activity by different demographic subgroups but lacked validation data for those algorithms across demographic subgroups [25,36-38]. To our knowledge, this is one of the first studies to show a proper validation approach to develop and test a generalizable algorithm across demographic subgroups where algorithm output could have differed by subgroup.

In addition, our approach highlights several points of interest when developing validation methodologies for this type of algorithm. The increased sample sizes and variability in data quality achieved by combining 2 distinct data sets enabled deeper characterization of the algorithm's performance. One of our studies generated data sets where truth labels were of high quality and accuracy but were collected from a study population limited in scope; the other study collected data from a large and demographically diverse cohort (albeit a somewhat engaged and self-selected participant group who volunteered and expressed interest in the PBHS and its health technology aspects), which allowed us to conduct subgroup analyses for both training and testing. Our results reinforce the well-established fact that modern machine-learning algorithms can sometimes perform well even when trained on a noisy data set [39]. This observation may be useful for researchers navigating study design decisions and tradeoffs, including

XSL•FO
RenderX

sample sizes and data labeling methods. For future research, determining the role of data quality factors in the development and characterization of this type of algorithm is an open issue [18].

Our approach to the generation of reference labels was pragmatic, using deployment-friendly ankle-worn devices or user-reported tags. Neither of these was as resource-demanding as other intensive approaches (ie, video monitoring), but generated information of sufficient quality to conduct our validation with relatively high time resolution (10-second epochs). Of note, the intrinsic nature of the 2 methods used for the generation of reference labels probably contributed to the noticeable difference in the proportions of "ambulatory" labels between the 2 studies (discussed in the Results), with the proportion observed in the pilot program study being the one closest to other literature reports [40].

When interpreting our results in the context of existing literature, it is worth noting that most validation studies for this type of algorithm have used step counts as the metric of interest [31,36,41-53], while ambulatory time (or a related metric) is the focus of a minority of reports [54,55]. In general, considering the close correlation between step count and ambulatory time, the performance of our algorithm could be placed on par with other algorithms, yet detailed side-by-side appraisals of results remain challenging; this research field is in need of standardization [19,56,57].

This study also had limitations. First, in principle, the StepWatch readouts used as ground truth may not have provided perfect accuracy, even though there is extensive literature supporting the use of StepWatch as a reference device [31,50,51,56,58-60]. Second, we observed fluctuations in the ambulatory status classification algorithm performance based on daily ambulatory time; this fluctuation was present when the algorithm detected 10-second epochs as ambulatory (or not) and was also manifested in the daily aggregates of ambulatory time. While this trend (shown in Figure 5) may have been driven, partially, by outlying data points with high step counts in our sample, which would be of little relevance in hypothetical clinical scenarios, it may also have been due to low-step periods containing mixed activities in which walking was not the only or dominant source of hand motion. In addition, while the cutoff used to read the StepWatch ambulatory classification relied on existing literature [61], it may not be perfect in itself. In this regard, it could be reassuring that the algorithm handled epochs with step counts between 4 and 8 as a continuum, as this is possibly reflective of the complexities of organic movement.

In sum, we have developed an accurate algorithm for the detection of the ambulatory status of users of a wrist-worn device in a free-living, real-world setting; the output is generalizable across several user demographic characteristics. The characterization of this algorithm was conducted in 2 distinct data sets, which lends credibility to the robustness and applicability of the performance results obtained in this study and illustrates the advantages of similar approaches to future research in this field.

Acknowledgments

The authors wish to thank David Andresen, Anthony Chan, Chen Chen, Robin Lin, Stephen Lanham, David Miller, Shannon Fong, and the Project Baseline Health Study team and participants for their contributions to this work. The authors also wish to acknowledge writing and editing support from Julia Saiz (Verily). This study was sponsored by Verily Life Sciences, which was responsible for data collection. Authors employed by Verily Life Sciences had access to the raw study data in full. All authors were responsible for interpretation of results and writing and review of the manuscript; all authors approved the final manuscript for submission.

Data Availability

The deidentified PBHS data corresponding to this study are available upon request for the purpose of examining their reproducibility. Interested investigators should direct requests to jsaiz@verily.com. Requests are subject to approval by the Project Baseline Health Study governance. Data from the pilot program are not available due to the nature of this program. Participants in this program did not consent for their data to be shared publicly.

Authors' Contributions

ER and RK contributed to study concept and design; ER contributed to data collection; and SP, MB, SS, ER, and RK contributed to data analysis and interpretation.

Conflicts of Interest

SP, MB, ER, SS, and RK report employment and equity ownership in Verily Life Sciences. JD is a scientific advisor to Veri, Inc.

Multimedia Appendix 1

Supplementary materials.
[DOCX File , 61 KB-Multimedia Appendix 1]

XSL•FO
RenderX

References

1. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, et al. The Physical Activity Guidelines for Americans. JAMA 2018 Nov 20;320(19):2020-2028 [FREE Full text] [doi: 10.1001/jama.2018.14854] [Medline: 30418471]
2. Katzmarzyk P, Powell KE, Jakicic JM, Troiano RP, Piercy K, Tennant B, 2018 Physical Activity Guidelines Advisory Committee. Sedentary Behavior and Health: Update from the 2018 Physical Activity Guidelines Advisory Committee. Med Sci Sports Exerc 2019 Jun;51(6):1227-1241 [FREE Full text] [doi: 10.1249/MSS.0000000000001935] [Medline: 31095080]
3. Johansen KL, Kaysen GA, Dalrymple LS, Grimes BA, Glidden DV, Anand S, et al. Association of physical activity with survival among ambulatory patients on dialysis: the Comprehensive Dialysis Study. Clin J Am Soc Nephrol 2013 Feb;8(2):248-253 [FREE Full text] [doi: 10.2215/CJN.08560812] [Medline: 23124787]
4. Waschki B, Kirsten A, Holz O, Müller KC, Meyer T, Watz H, et al. Physical activity is the strongest predictor of all-cause mortality in patients with COPD: a prospective cohort study. Chest 2011 Aug;140(2):331-342. [doi: 10.1378/chest.10-2521] [Medline: 21273294]
5. Walsh JT, Charlesworth A, Andrews R, Hawkins M, Cowley AJ. Relation of daily activity levels in patients with chronic heart failure to long-term prognosis. Am J Cardiol 1997 May 15;79(10):1364-1369. [doi: 10.1016/s0002-9149(97)00141-0] [Medline: 9165159]
6. Garcia-Aymerich J, Lange P, Benet M, Schnohr P, Antó JM. Regular physical activity reduces hospital admission and mortality in chronic obstructive pulmonary disease: a population based cohort study. Thorax 2006 Sep;61(9):772-778 [FREE Full text] [doi: 10.1136/thx.2006.060145] [Medline: 16738033]
7. Garcia-Rio F, Rojo B, Casitas R, Lores V, Madero R, Romero D, et al. Prognostic value of the objective measurement of daily physical activity in patients with COPD. Chest 2012 Aug;142(2):338-346. [doi: 10.1378/chest.11-2014] [Medline: 22281798]
8. Alahmari AD, Patel AR, Kowlessar BS, Mackay AJ, Singh R, Wedzicha JA, et al. Daily activity during stability and exacerbation of chronic obstructive pulmonary disease. BMC Pulm Med 2014 Jun 02;14:98 [FREE Full text] [doi: 10.1186/1471-2466-14-98] [Medline: 24885188]
9. Hayata A, Minakata Y, Matsunaga K, Nakanishi M, Yamamoto N. Differences in physical activity according to mMRC grade in patients with COPD. Int J Chron Obstruct Pulmon Dis 2016;11:2203-2208 [FREE Full text] [doi: 10.2147/COPD.S109694] [Medline: 27695306]
10. Demeyer H, Gimeno-Santos E, Rabinovich RA, Hornikx M, Louvaris Z, de Boer WI, PROactive consortium. Physical activity characteristics across GOLD quadrants depend on the questionnaire used. PLoS One 2016;11(3):e0151255 [FREE Full text] [doi: 10.1371/journal.pone.0151255] [Medline: 26974332]
11. Chow E, Abdolell M, Panzarella T, Harris K, Bezjak A, Warde P, et al. Predictive model for survival in patients with advanced cancer. J Clin Oncol 2008 Dec 20;26(36):5863-5869. [doi: 10.1200/JCO.2008.17.1363] [Medline: 19018082]
12. Maltoni M, Caraceni A, Brunelli C, Broeckaert B, Christakis N, Eychmueller S, Steering Committee of the European Association for Palliative Care. Prognostic factors in advanced cancer patients: evidence-based clinical recommendations--a study by the Steering Committee of the European Association for Palliative Care. J Clin Oncol 2005 Sep 01;23(25):6240-6248. [doi: 10.1200/JCO.2005.06.866] [Medline: 16135490]
13. Ballard-Barbash R, Friedenreich CM, Courneya KS, Siddiqi SM, McTiernan A, Alfano CM. Physical activity, biomarkers, and disease outcomes in cancer survivors: a systematic review. J Natl Cancer Inst 2012 Jun 06;104(11):815-840 [FREE Full text] [doi: 10.1093/jnci/djs207] [Medline: 22570317]
14. Pitta F, Troosters T, Probst VS, Spruit MA, Decramer M, Gosselink R. Quantifying physical activity in daily life with questionnaires and motion sensors in COPD. Eur Respir J 2006 May;27(5):1040-1055 [FREE Full text] [doi: 10.1183/09031936.06.00064105] [Medline: 16707399]
15. Fiedler J, Eckert T, Burchartz A, Woll A, Wunsch K. Comparison of self-reported and device-based measured physical activity using measures of stability, reliability, and validity in adults and children. Sensors (Basel) 2021 Apr 10;21(8):2672 [FREE Full text] [doi: 10.3390/s21082672] [Medline: 33920145]
16. Skender S, Ose J, Chang-Claude J, Paskow M, Brühmann B, Siegel EM, et al. Accelerometry and physical activity questionnaires - a systematic review. BMC Public Health 2016 Jun 16;16(1):515 [FREE Full text] [doi: 10.1186/s12889-016-3172-0] [Medline: 27306667]
17. Düking P, Fuss FK, Holmberg H, Sperlich B. Recommendations for assessment of the reliability, sensitivity, and validity of data provided by wearable sensors designed for monitoring physical activity. JMIR Mhealth Uhealth 2018 Apr 30;6(4):e102 [FREE Full text] [doi: 10.2196/mhealth.9341] [Medline: 29712629]
18. Poli A, Cosoli G, Scalise L, Spinsante S. Impact of Wearable Measurement Properties and Data Quality on ADLs Classification Accuracy. IEEE Sensors J 2021 Jul 1;21(13):14221-14231. [doi: 10.1109/jsen.2020.3009368]
19. Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). NPJ Digit Med 2020;3:55 [FREE Full text] [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]

XSL•FO
RenderX

20. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020 Sep 08;8(9):e18694 [FREE Full text] [doi: 10.2196/18694] [Medline: 32897239]
21. Moore C, McCullough AK, Aguiar EJ, Ducharme SW, Tudor-Locke C. Toward harmonized treadmill-based validation of step-counting wearable technologies: a scoping review. J Phys Act Health 2020 Jul 11:1-13 [FREE Full text] [doi: 10.1123/jpah.2019-0205] [Medline: 32652514]
22. Knaier R, Höchsmann C, Infanger D, Hinrichs T, Schmidt-Trucksäss A. Validation of automatic wear-time detection algorithms in a free-living setting of wrist-worn and hip-worn ActiGraph GT3X. BMC Public Health 2019 Feb 28;19(1):244 [FREE Full text] [doi: 10.1186/s12889-019-6568-9] [Medline: 30819148]
23. Poli A, Spinsante S, Nugent C, Cleland I. Improving the collection and understanding the quality of datasets for the aim of human activity recognition. In: Chen F, García-Betances R, Chen L, Cabrera-Umpiérrez M, Nugent C, editors. Smart Assisted Living. Cham, Switzerland: Springer; 2020.
24. Prince SA, Adamo KB, Hamel M, Hardt J, Connor Gorber S, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008 Nov 06;5:56 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
25. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. Sci Rep 2018 May 21;8(1):7961 [FREE Full text] [doi: 10.1038/s41598-018-26174-1] [Medline: 29784928]
26. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al. Large scale population assessment of physical activity using wrist worn accelerometers: the UK Biobank Study. PLoS One 2017;12(2):e0169649 [FREE Full text] [doi: 10.1371/journal.pone.0169649] [Medline: 28146576]
27. Arges K, Assimes T, Bajaj V, Balu S, Bashir MR, Beskow L, et al. The Project Baseline Health Study: a step towards a broader mission to map human health. NPJ Digit Med 2020;3:84 [FREE Full text] [doi: 10.1038/s41746-020-0290-y] [Medline: 32550652]
28. Bloem B, Marks WJ, Silva de Lima AL, Kuijf ML, van Laar T, Jacobs BPF, et al. The Personalized Parkinson Project: examining disease progression through broad biomarkers in early Parkinson's disease. BMC Neurol 2019 Jul 17;19(1):160 [FREE Full text] [doi: 10.1186/s12883-019-1394-3] [Medline: 31315608]
29. Burq M, Rainaldi E, Ho KC, Chen C, Bloem BR, Evers LJW, et al. Virtual exam for Parkinson's disease enables frequent and reliable remote measurements of motor function. NPJ Digit Med 2022 May 23;5(1):65 [FREE Full text] [doi: 10.1038/s41746-022-00607-8] [Medline: 35606508]
30. McLean SA, Ressler K, Koenen KC, Neylan T, Germine L, Jovanovic T, et al. The AURORA Study: a longitudinal, multimodal library of brain biology and function after traumatic stress exposure. Mol Psychiatry 2020 Feb;25(2):283-296 [FREE Full text] [doi: 10.1038/s41380-019-0581-3] [Medline: 31745239]
31. Toth L, Park S, Springer CM, Feyerabend MD, Steeves JA, Bassett DR. Video-recorded validation of wearable step counters under free-living conditions. Med Sci Sports Exerc 2018 Jun;50(6):1315-1322. [doi: 10.1249/MSS.0000000000001569] [Medline: 29381649]
32. Cho J. Current status and prospects of health-related sensing technology in wearable devices. J Healthc Eng 2019;2019:3924508 [FREE Full text] [doi: 10.1155/2019/3924508] [Medline: 31316740]
33. Keogh A, Dorn JF, Walsh L, Calvo F, Caulfield B. Comparing the usability and acceptability of wearable sensors among older Irish adults in a real-world context: observational study. JMIR Mhealth Uhealth 2020 Apr 20;8(4):e15704 [FREE Full text] [doi: 10.2196/15704] [Medline: 32310149]
34. Huberty J, Ehlers DK, Kurka J, Ainsworth B, Buman M. Feasibility of three wearable sensors for 24 hour monitoring in middle-aged women. BMC Womens Health 2015 Jul 30;15:55 [FREE Full text] [doi: 10.1186/s12905-015-0212-3] [Medline: 26223521]
35. Chu AHY, Ng SHX, Paknezhad M, Gauterin A, Koh D, Brown MS, et al. Comparison of wrist-worn Fitbit Flex and waist-worn ActiGraph for measuring steps in free-living adults. PLoS One 2017;12(2):e0172535 [FREE Full text] [doi: 10.1371/journal.pone.0172535] [Medline: 28234953]
36. Golbus JR, Pescatore NA, Nallamothu BK, Shah N, Kheterpal S. Lancet Digit Health 2021 Nov;3(11):e707-e715 [FREE Full text] [doi: 10.1016/S2589-7500(21)00138-2] [Medline: 34711377]
37. Sabia S, van Hees VT, Shipley MJ, Trenell MI, Hagger-Johnson G, Elbaz A, et al. Association between questionnaire- and accelerometer-assessed physical activity: the role of sociodemographic factors. Am J Epidemiol 2014 Mar 15;179(6):781-790 [FREE Full text] [doi: 10.1093/aje/kwt330] [Medline: 24500862]
38. Gupta S, Gupta A. Dealing with noise problem in machine learning data-sets: a systematic review. Procedia Computer Science 2019;161:466-474 [FREE Full text] [doi: 10.1016/j.procs.2019.11.146]
39. Wallmann-Sperlich B, Chau JY, Froboese I. Self-reported actual and desired proportion of sitting, standing, walking and physically demanding tasks of office employees in the workplace setting: do they fit together? BMC Res Notes 2017 Nov 17;10(1):504 [FREE Full text] [doi: 10.1186/s13104-017-2829-9] [Medline: 29145883]

XSL•FO
RenderX

40. Giansanti D, Tiberi Y, Silvestri G, Maccioni G. New wearable system for step-counting telemonitoring and telerehabilitation based on the Codivilla spring. Telemed J E Health 2008 Dec;14(10):1096-1100. [doi: 10.1089/tmj.2008.0035] [Medline: 19119833]
41. Magistro D, Brustio PR, Ivaldi M, Esliger DW, Zecca M, Rainoldi A, et al. Validation of the ADAMO Care Watch for step counting in older adults. PLoS One 2018;13(2):e0190753 [FREE Full text] [doi: 10.1371/journal.pone.0190753] [Medline: 29425196]
42. Germini F, Noronha N, Borg Debono V, Abraham Philip B, Pete D, Navarro T, et al. Accuracy and acceptability of wrist-wearable activity-tracking devices: systematic review of the literature. J Med Internet Res 2022 Jan 21;24(1):e30791 [FREE Full text] [doi: 10.2196/30791] [Medline: 35060915]
43. Chen M, Kuo CC, Pellegrini CA, Hsu MJ. Accuracy of wristband activity monitors during ambulation and activities. Med Sci Sports Exerc 2016 Oct;48(10):1942-1949. [doi: 10.1249/MSS.0000000000000984] [Medline: 27183123]
44. Park S, Marcotte RT, Toth LP, Paulus P, Lauricella LP, Kim AH, et al. Free-living validation and harmonization of 10 wearable step count monitors. Transl J ACSM 2021 Sep 3;6(4):e000172 [FREE Full text] [doi: 10.1249/tjx.0000000000000172]
45. Polese JC, E Faria GS, Ribeiro-Samora GA, Lima LP, Coelho de Morais Faria CD, Scianni AA, et al. Google fit smartphone application or Gt3X Actigraph: Which is better for detecting the stepping activity of individuals with stroke? A validity study. J Bodyw Mov Ther 2019 Jul;23(3):461-465. [doi: 10.1016/j.jbmt.2019.01.011] [Medline: 31563356]
46. Floegel TA, Florez-Pregonero A, Hekler EB, Buman MP. Validation of consumer-based hip and wrist activity monitors in older adults with varied ambulatory abilities. J Gerontol A Biol Sci Med Sci 2017 Feb;72(2):229-236 [FREE Full text] [doi: 10.1093/gerona/glw098] [Medline: 27257217]
47. Toth LP. Validity of activity tracker step counts during walking, running, and activities of daily living. Transl J Am Coll Sports Med 2018;3:52-59 [FREE Full text] [doi: 10.1249/TJX.0000000000000057]
48. Korpan S, Schafer JL, Wilson KC, Webber SC. Effect of ActiGraph GT3X+ position and algorithm choice on step count accuracy in older adults. J Aging Phys Act 2015 Jul;23(3):377-382. [doi: 10.1123/japa.2014-0033] [Medline: 25102469]
49. Webber SC, St John PD. Comparison of ActiGraph GT3X+ and StepWatch step count accuracy in geriatric rehabilitation patients. J Aging Phys Act 2016 Jul;24(3):451-458. [doi: 10.1123/japa.2015-0234] [Medline: 26751505]
50. Feito Y, Bassett DR, Thompson DL. Evaluation of activity monitors in controlled and free-living environments. Med Sci Sports Exerc 2012 Apr;44(4):733-741. [doi: 10.1249/MSS.0b013e3182351913] [Medline: 21904249]
51. Bender CG, Hoffstot JC, Combs BT, Hooshangi S, Cappos J. Measuring the fitness of fitness trackers. 2017 Presented at: 2017 IEEE Sensors Applications Symposium (SAS); March 13-15, 2017; Glassboro, NJ p. 1-6. [doi: 10.1109/sas.2017.7894077]
52. Hickey A, John D, Sasaki JE, Mavilia M, Freedson P. Validity of activity monitor step detection is related to movement patterns. J Phys Act Health 2016 Feb;13(2):145-153. [doi: 10.1123/jpah.2015-0203] [Medline: 26107045]
53. Bai Y, Tompkins C, Gell N, Dione D, Zhang T, Byun W. Comprehensive comparison of Apple Watch and Fitbit monitors in a free-living setting. PLoS One 2021;16(5):e0251975 [FREE Full text] [doi: 10.1371/journal.pone.0251975] [Medline: 34038458]
54. Burnham J, Lu C, Yaeger LH, Bailey TC, Kollef MH. Using wearable technology to predict health outcomes: a literature review. J Am Med Inform Assoc 2018 Sep 01;25(9):1221-1227 [FREE Full text] [doi: 10.1093/jamia/ocy082] [Medline: 29982520]
55. Walmsley R, Chan S, Smith-Byrne K, Ramakrishnan R, Woodward M, Rahimi K, et al. Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. Br J Sports Med 2021 Sep 06;56(18):1008-1017 [FREE Full text] [doi: 10.1136/bjsports-2021-104050] [Medline: 34489241]
56. Hergenroeder AL, Barone Gibbs B, Kotlarczyk MP, Kowalsky RJ, Perera S, Brach JS. Accuracy of Objective Physical Activity Monitors in Measuring Steps in Older Adults. Gerontol Geriatr Med 2018;4:2333721418781126 [FREE Full text] [doi: 10.1177/2333721418781126] [Medline: 29977979]
57. Bent B, Wang K, Grzesiak E, Jiang C, Qi Y, Jiang Y, et al. The digital biomarker discovery pipeline: An open-source software platform for the development of digital biomarkers using mHealth and wearables data. J Clin Transl Sci 2020 Jul 14;5(1):e19 [FREE Full text] [doi: 10.1017/cts.2020.511] [Medline: 33948242]
58. Treacy D, Hassett L, Schurr K, Chagpar S, Paul SS, Sherrington C. Validity of Different Activity Monitors to Count Steps in an Inpatient Rehabilitation Setting. Phys Ther 2017 May 01;97(5):581-588. [doi: 10.1093/ptj/pzx010] [Medline: 28339904]
59. Moy ML, Danilack VA, Weston NA, Garshick E. Daily step counts in a US cohort with COPD. Respir Med 2012 Jul;106(7):962-969 [FREE Full text] [doi: 10.1016/j.rmed.2012.03.016] [Medline: 22521225]
60. Schmidt A, Pennypacker ML, Thrush AH, Leiper CI, Craik RL. Validity of the StepWatch Step Activity Monitor: preliminary findings for use in persons with Parkinson disease and multiple sclerosis. J Geriatr Phys Ther 2011;34(1):41-45. [doi: 10.1519/JPT.0b013e31820aa921] [Medline: 21937891]
61. Kluge F, Del Din S, Cereatti A, Gaßner H, Hansen C, Helbostad JL, Mobilise-D consortium. Consensus based framework for digital mobility monitoring. PLoS One 2021;16(8):e0256541 [FREE Full text] [doi: 10.1371/journal.pone.0256541] [Medline: 34415959]

XSL•FO
RenderX

Abbreviations

AUC: area under the receiver operating characteristic curve
IRB: institutional review board
MAE: mean absolute error
MAPE: mean absolute percentage error
PBHS: Project Baseline Health Study
PPV: positive predictive value
PRC: precision-recall curve
QC: quality control
ROC: receiver operating characteristic

Edited by G Eysenbach; submitted 21.10.22; peer-reviewed by M Kraus, A Gangadhara Rao; comments to author 24.11.22; revised version received 05.12.22; accepted 19.01.23; published 07.03.23

Please cite as:
Popham S, Burq M, Rainaldi EE, Shin S, Dunn J, Kapur R
An Algorithm to Classify Real-World Ambulatory Status From a Wearable Device Using Multimodal and Demographically Diverse Data: Validation Study
JMIR Biomed Eng 2023;8:e43726
URL: https://biomedeng.jmir.org/2023/1/e43726
doi: 10.2196/43726
PMID:

©Sara Popham, Maximilien Burq, Erin E Rainaldi, Sooyoon Shin, Jessilyn Dunn, Ritu Kapur. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 07.03.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX