SUPPLEMENTAL MATERIAL

Data S1.

SUPPLEMENTAL METHODS

Sample Size Estimate Calculations

A threshold of 67% for per-interval sensitivity was chosen as continuous monitoring provides multiple opportunities to detect periods of suspected AF. For example, If AF occurs over two 15-min intervals then the effective sensitivity is 89%, which increases to 96% when there are 3 AF-positive 15-minute intervals (or a single 45-min episode). A minimum of 90% was set for per-interval specificity as that would yield approximately 4 false-positive notifications per day, which would be 4 additional ECG prompts per day. This was viewed as an acceptable upper limit for a prescription-only device in subjects that have been diagnosed with or are vulnerable to developing AF. These thresholds served as the basis for our primary endpoints.

Sample size estimates are based on the difference between the primary endpoint thresholds described above (67% sensitivity and 90% specificity) and the expected performance of our test device (80% sensitivity, 95% Specificity), where differences in sensitivity and specificity are measured at the interval-level and intervals are clustered by subject. We assumed 470 intervals captured per subject containing sufficient quality PPG data based on a projected average wear time per subject of 14 hours/day (4 intervals/hour x 14 hours/day x 14 days = 784 intervals. We assumed that 60% of intervals will have sufficient quality PPG data resulting in 470 intervals (784 x 0.6). An average of 17% of captured intervals per subject were projected to be reference positive, based on previous studies of subjects with AF. Prior data was also used to estimate an intracluster correlation coefficient (ICC) of r = 0.5, which is a measuring of AF occurrences within-subject vs. across-subject.   Finally, we assumed 10% loss of subjects that results in unevaluable data, as a result of loss to follow-up, non-compliance or device error.

The effective sample size is based on a formula[30] that penalizes the nominal sample size calculation on an estimated design-effect:

n_c = n * [1 + (s - 1) * r] where,

- n_c (effective sample size) = the number of intervals required for the study accounting for clustering
- n (nominal sample size) = the number of positive or negative intervals that would be required if these intervals were truly independent based on providing 80% Power (at one-sided alpha = 2.5%)
- [1 + (s - 1) * r] represents the design effect/variance inflation factor, where
- s = average cluster size. This is the average number of ref+ or ref- intervals per subject that will have sufficient quality test/reference data.
- r = the proportion of total variance explained by across-cluster (e.g. subject) variance. Measured using prior clinical data.

Allowing for multiple dependent intervals per subject and based on the assumptions above (i.e. r = 0.50, AF burden = 17%), approximately 110 evaluable subjects are required to provide 80% power to detect a difference in specificity of 5%. This same number of evaluable subjects would provide >90% power to detect a difference in sensitivity of 13%. Accounting for study loss, up to 140 subjects were estimated to be needed to obtain this number of evaluable subjects.

Logistic Mixed-Effect Regression Models to Estimate Accuracy

Estimates of accuracy were calculated across intervals, while accounting for clustering of intervals within-subject. Logistic mixed-effect regression models, with a random effect added to account for the effect of clustering, were used to estimate sensitivity, specificity, PPV, and NPV. The correlation of repeated measurements on the same individual was specified by including the $\boldsymbol{\gamma}$ vector of random effects and then specifying the structure of the variance-covariance matrix of the random effects. We used an unstructured covariance, which estimates unique correlations for each pair of intervals. However, no term was included that accounts for correlation of intervals with lagged versions of those same intervals (autocorrelation).

Confidence Interval Estimation

For all estimates of accuracy at the interval level, 95% Confidence Intervals (CIs) were provided using bootstrap methods (‘cluster’ or ‘block’ bootstrap) where the correlation structure is preserved, taking into account the fact that repeat intervals are nested within subjects (i.e. individual). Standard bootstrapping procedures require identically distributed and independent responses, which is not the case with such nested data.

Cluster bootstrapping, modifies the standard bootstrapping procedure with regard to the resampling process as follows:

- Define: J = cluster unit = subject, where multiple observation units (intervals) may be observed
- The sampling is based on the total number of J clusters.
    - The first step is to randomly select J number of clusters with replacement
    - For each cluster selected (with some clusters selected more than once and others not selected at all), all observations within that cluster are selected. Original cluster sizes are maintained.
    - Sensitivity or specificity (θ) are computed using the bootstrapped sample and the process is repeated B number of times. Our analysis used B = 10,000.
- Non-parametric confidence intervals were derived based on the 2.5% and 97.5% quantiles of the resulting bootstrap distribution (i.e. of the ordered distribution of θs).
- Point estimates for sensitivity and specificity were based on model estimates of the proportions, whereas 95% CIs were based on the bootstrap distribution.

Clopper-Pearson intervals were used only as noted when cluster-bootstrap intervals could not be calculated due to low error counts (i.e. non-convergence). In such cases within-subject correlations are not accounted for. In all other cases cluster-bootstrap intervals are used.

Model Architecture

The on-watch AF detection algorithm is a Convolution Neural Network (CNN) based model that learns features from PPG signal. Compared to using detected beats which are susceptible to noise or motion from PPG, these extracted features are more likely to maintain signal morphology and yield robust prediction results even when signals are noisy. The entire model for on-board prompting is designed compactly such that it can be integrated into the Study Watch firmware, consisting of 7 convolutional layers including an additional fully connected output layer.

The AF burden estimation algorithm is based on a patented encoder-decoder scheme. A CNN-based model serves as an encoder on-watch, and a residual neural network (ResNet) as a decoder that is run server-side instead of the firmware, using PPG features that are outputted from the encoder on the firmware and transmitted from the device. This approach allows for a computationally lighter 3-layer model to run on the firmware, while a more complex model is run server-side, incorporating 8 residual blocks with two convolutional layers each, totalling 17 convolutional layers including an additional fully connected layer.

Algorithm Development

Training

To train the algorithms, we created a dataset of over 400,000 30-second PPG strips by simulating synthetic PPG waveforms from 30-second modified lead II ECG strips enriched with a variety of arrhythmia from the Zio population. Each of these ECG strips was reviewed by Certified Cardiographic Technicians (CCTs) to obtain the corresponding rhythm labels. Synthetic noise was also added to simulate electrical noise and motion artifacts present in real PPG signals, augmenting the dataset. Each PPG strip was preprocessed (i.e., resampled, filtered, and normalized) and fed as input for model training.

Tuning

The tuning process employed a bootstrapping approach to ensure that the algorithms are generalizable for use with the intended population. Both algorithms aggregate prediction results over an interval of time and output AF and unanalyzable prediction probabilities for on-wrist intervals. These probabilities are compared against a threshold to determine if the interval will be labeled unanalyzable, AF, or not AF. The thresholds were determined using ambulatory PPG and corresponding ECG-based rhythm labels from Study Watch Atrial Fibrillation (AF) Detection Investigation (SWAFDI; NCT04074434), a previously run study independent from the current evaluation study "Study Watch AF Detection At Home" (NCT04546763).

Table S1. PPG interbeat interval variability comparison for false positive and true negative results (Two subjects with highest counts vs. remaining false positive individuals).

PPG IBI intervals	Subject #1 (1,148 intervals)	Subject #2 (951 intervals)	All other 43 subjects with FPs (33,834 intervals)	All FPs (1,599 FPs)	All TNs (8,2972 TNs)	All TPs (7,000 TPs)
SDNN (ms)	191.0 ± 49.2	157.6 ± 61.6	148.8 ± 75.9	176.90 ± 53.90	130.45 ± 72.33	205.09 ± 45.49
RMSSD (ms)	261.8 ± 60.3	203.3 ± 87.2	191.4 ± 110.2	241.25 ± 71.49	163.78 ± 100.73	270.74 ± 58.33

SDNN: Standard Deviation of NN intervals; RMSSD: stands for Root Mean Square of Successive Differences; SD: standard deviation; IBI: Interbeat interval variability (mean ± SD); FP: false positive; TN: true negative; TP: true positive

Table S2. Sensitivity analysis to assess the impact of un-analyzable and off-wrist test intervals on performance.

	Assuming un-analyzable + off-wrist test intervals are:			
	Test-Positive	**Test-Negative**	**Matches Reference**	**Does Not Match Reference**
Per interval Sensitivity (95% CI)	97.2% [94.8%, 98.5%]	69.2% [60.5%, 76.5%]	97.2% [94.8%, 98.5%]	69.2% [60.5%, 76.5%]
Per Interval Specificity (95% CI)	68.8% [65.8%, 71.7%]	98.7% [98.1%, 99.3%]	98.7% [98.1%, 99.3%]	68.8% [65.8%, 71.7%]

Table S3. Unanalyzable intervals by activity level.

Activity Level	# of intervals with Ziopatch data	# of unanalyzable intervals	% of unanalyzable intervals per activity level	% of all unanalyzable intervals (27,115 intervals)
Light	105,146	17,078	16.2%	63.0%
Moderate	12,043	8,457	70.2%	31.2%
Vigorous	1,783	1,580	88.6%	5.8%

Table S4. Unanalyzable intervals by arm hair index.

Arm Hair Index	# of unanalyzable intervals	% of unanalyzable intervals within the subgroup	% of all unanalyzable intervals (27,115 intervals)
1 (N = 47)	11,379	23.0%	42.0%
2 (N = 32)	7,473	22.7%	28.2%
3 (N = 29)	7,649	23.1%	27.6%
4 (N = 3)	614	17.9%	2.2%

Table S5. Unanalyzable intervals by Fitzpatrick skin tone.

Skin Tone	# of unanalyzable intervals	% of unanalyzable intervals within the subgroup	% of all unanalyzable intervals (27,115 intervals)
I (N = 10)	1,696	14.3%	6.2%
II (N = 17)	3,600	18.8%	13.3%
III (N = 42)	8,610	20.0%	31.8%
IV (N = 26)	6,257	22.6%	23.1%
V/VI (N = 16)	6,952	40.2%	25.6%

Figure S1. Example of false positive PPG interval with high IBI. The filtered PPG and patch ECG waveforms from a 15-min interval from Subject # 1 in Table S1 (one minute of data per row). The SDNN from PPG-derived interbeat intervals (IBIs) is 166.9 ms and RMSSD is 264.5 ms.

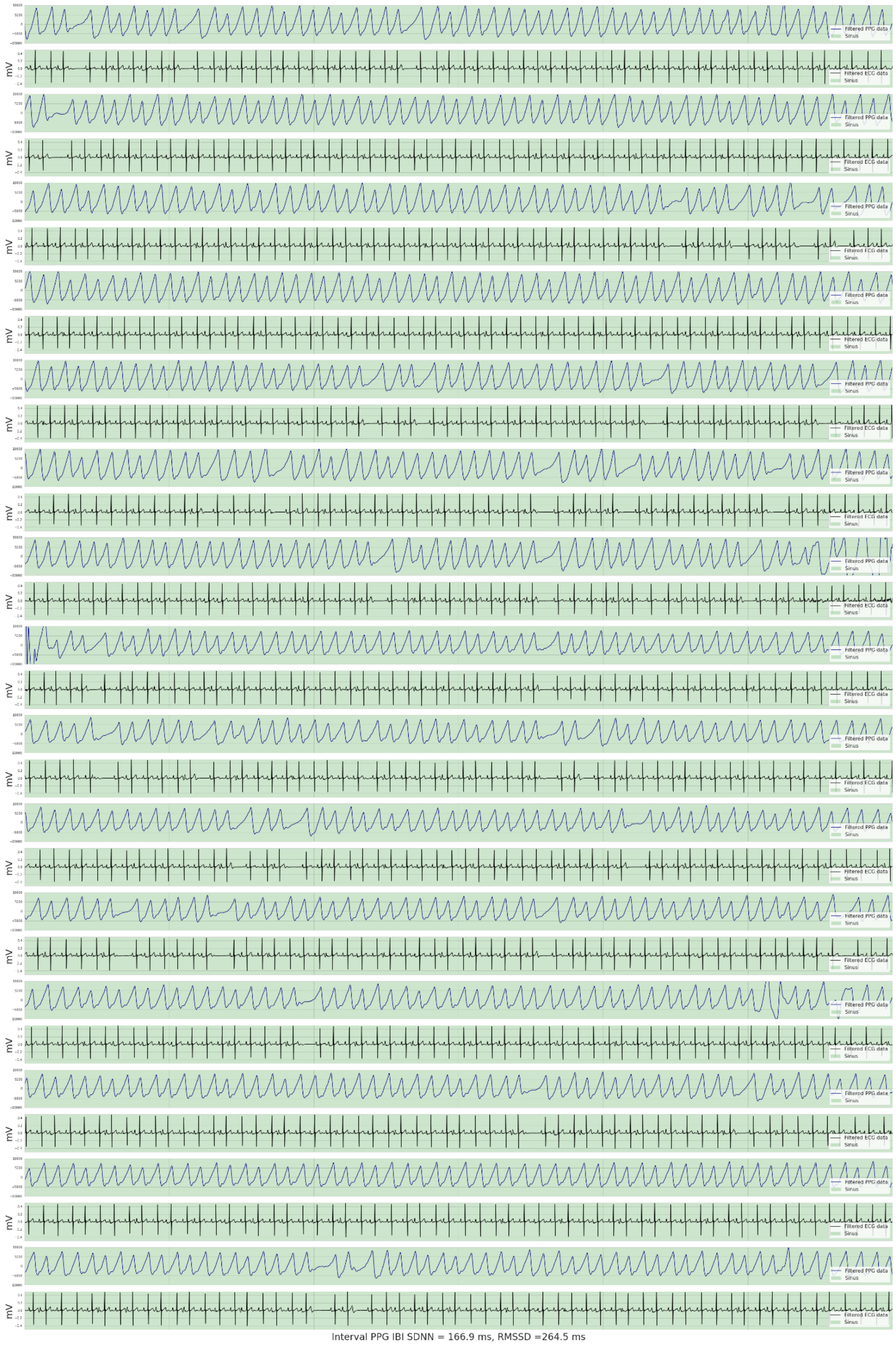


Figure S2. Example of false negative PPG interval with low AF burden. The filtered PPG and patch ECG waveforms from a 15-min interval with aggregated AF duration of 103 seconds highlighted in red. There is one minute of data per row. The SDNN from PPG-derived interbeat intervals (IBIs) is 114.7 ms and RMSSD is 88.7 ms.

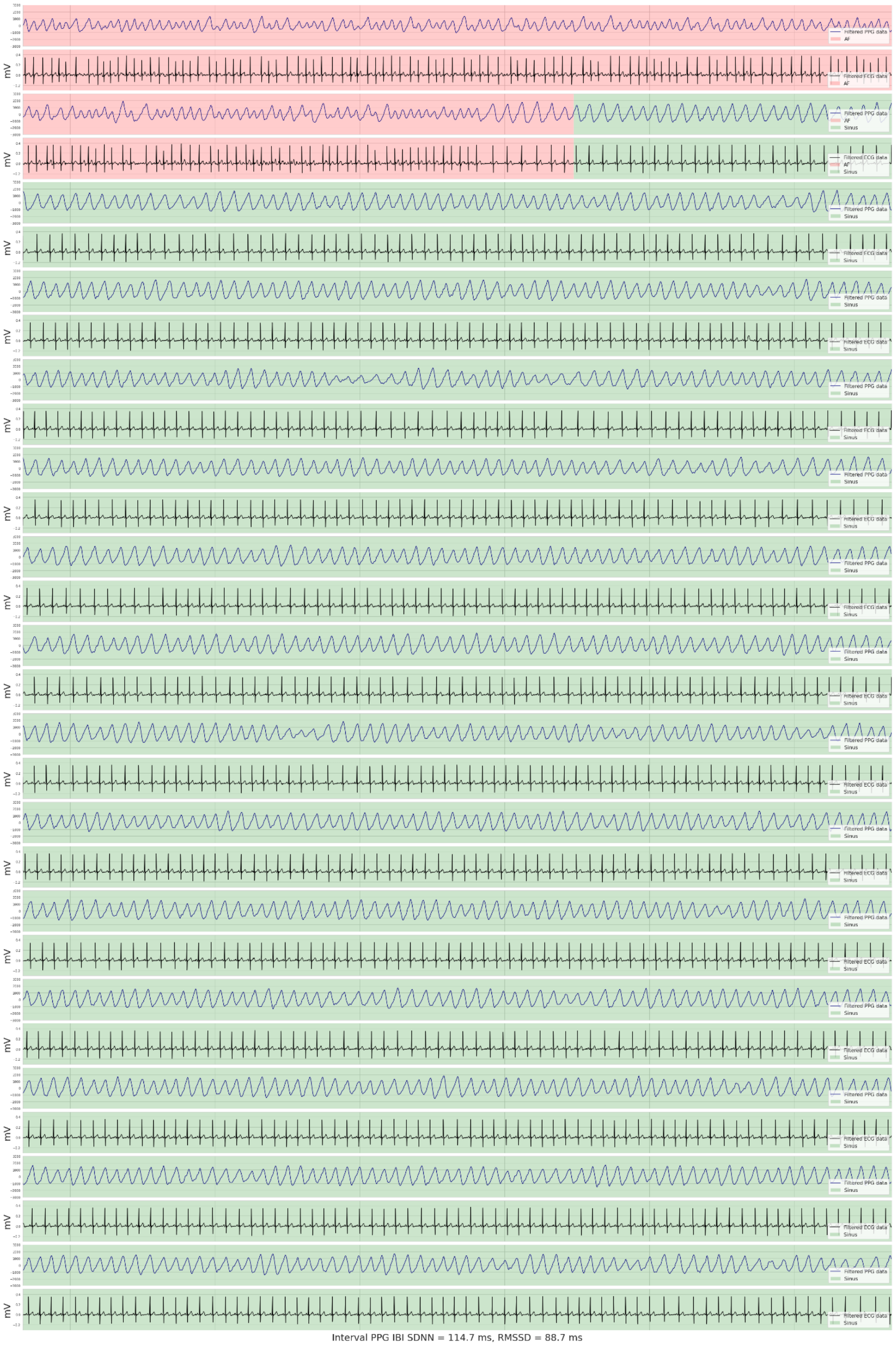
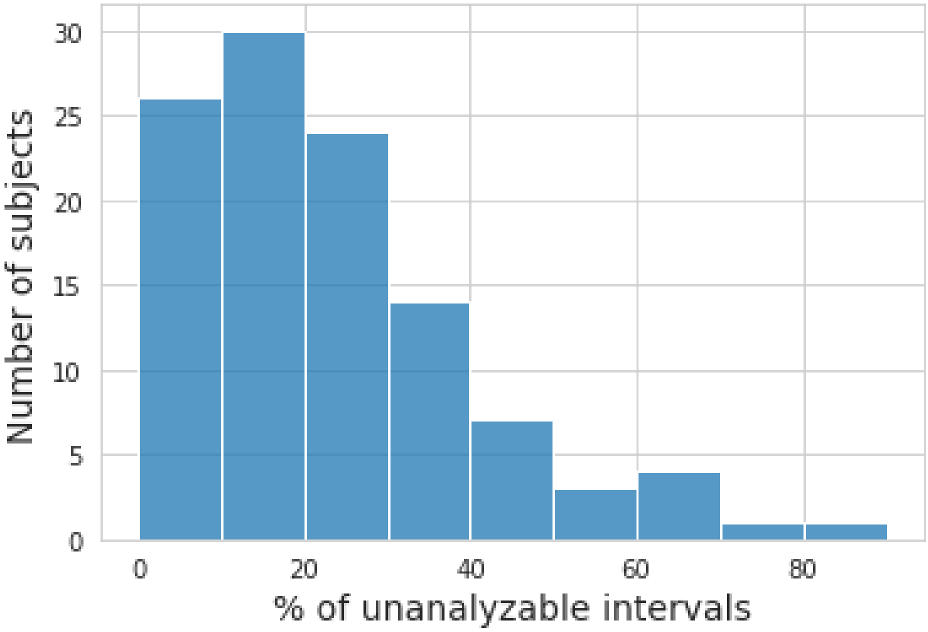


Figure S3. Distribution of patients across percentiles of un-analyzable intervals.


Number of subjects
% of unanalyzable intervals