CoughMonitor Pilot Validation: Data Analysis and Presentation

Hyfe Data Science

2023-03-01

1 Introduction

The Hyfe "CoughMonitor" is a wrist-worn device that captures sound through its microphone and processes that sound on device to detect and quantify coughing over time; the device ingests continuous audio and outputs the timestamps of all of the coughs that it detects. To assess the CoughMonitor's accuracy, one must compare its output (when the device reports that coughing occurred) with objective ground truth (when trained human annotators report that coughing occurred). A perfectly accurate device would detect every human annotated cough as such and would not identify anything else as a cough.

This paper describes a small pilot study aimed at quantifying the accuracy of an earlier version of the Hyfe CoughMonitor and, in so doing, demonstrates the analytical approach and data presentation that will be employed in the subsequent FDA enabling study.

2 Methods

2.1 Clinical and Annotation

Over the course of February 2023, 4 patients were enrolled into a pilot trial to "run water through the pipes" of the program that will be used for the definitive validation of the final version of the Hyfe Cough Monitoring System (HCMS). Two of these patients were enrolled at the Oregon Health & Science University (OHSU) in the United States and two were enrolled at the University of Navarra in Spain. All 4 patients were instructed to carry out their usual activities of daily life while using the CoughMonitor for 24 continuous hours. In order to generate ground truth from continuous audio, each participant also wore (on their other wrist) an audio recording device, the "Validator", which captured and stored all 24 hours of their monitoring session. At bedtime, each patient removed the devices from their wrists and plugged them in at their bedside to charge during nighttime monitoring. Following the monitoring sessions, devices were returned to researchers, and the patients' data – timestamps of coughs from the CoughMonitor as well as continuous audio files from the Validator – were uploaded via wifi to cloud storage for analysis.

Prior to analysis, all 24 hours of monitoring data for each participant were sliced into 60-second segments. A team of trained annotators then listened to all of the segmented audio; two annotators listened independently to each 60-second segment. The annotators simultaneously heard the audio and watched visualizations of it (waveform and mel-spectrogram) in a custom-made web application that allowed them to listen to any segment, or any sub-section thereof, as many times as they wished. Using this application, annotators labeled exactly when each cough occurred in each 60-second segment. For research purposes, annotators tagged ambiguous cough like sounds as "Not sure" and tagged very faint cough like sounds as "Far".

Occasionally, annotators disagreed regarding exactly if/when a cough occurred in the same 60-second audio segment. These discrepant sounds were adjudicated by a third listener who has special expertise in cough annotation. Only cough sounds were included in the subsequent analysis. Consequently, ground truth consisted of nearby, unambiguous coughing events labeled by at least two independent individuals, with difficult sounds – those for which annotators disagreed or indicated uncertainty – adjudicated by an MD.

2.2 Coughs and cough-seconds

The labeling process established the ground truth against which the CoughMonitor's output was compared. Two basic units of analysis were used:

- *coughs*, individually time stamped by the human annotators and by the CoughMonitor, and
- *cough-seconds*, defined to be seconds containing at least one timestamped cough.

Cough-seconds are useful because coughs sometimes occur in rapid succession, with or without intervening inhalation, occasionally making it difficult to distinguish between the end of one cough and the beginning of the next cough during a rapid burst of 2 or 3 explosive sounds.

2.3 Correlation analysis

It is important to understand the sensitivity and specificity of diagnostic tests in which each assay can have profound clinical implications. However, unlike

diagnostic test results, an individual cough has no clinical value. What is important to patients, providers and researchers (and is the readout of the HCMS) is the cough rate which is best expressed as the number of coughs per hour. Thus, our preferred metric of accuracy is the correlation between the hourly cough rates determined by human annotators and by the HCMS.

Participants' coughs and cough-seconds, as determined by the ground truth annotations and by the CoughMonitor's output, were tabulated hourly and compared. Each participant therefore contributed two pairs of counts per hour of monitoring; for simplicity, we refer to these distinct endpoints simultaneously as "hourly cough counts."

If the CoughMonitor's output agreed perfectly with the ground truth annotations, then the paired hourly cough counts would all lie on the line $y = x$ (ground truth counts on the x-axis, CoughMonitor counts on the y-axis). Measures of the agreement with this ideal line provide three CoughMonitor accuracy metrics:

- the *Pearson correlation coefficient* between the paired hourly cough counts, and
- the *slope* and *intercept* of the regression line that summarizes the relationship between the paired hourly cough counts.

These linear analysis metrics were calculated for the entire cohort and separately for each individual participant; scatterplots display the hourly cough counts and the lines of best fit.

If the CoughMonitor's performance did not depend on cough frequency, then the percentage errors of its hourly cough counts, relative to the ground truth hourly cough counts, would vary uniformly about 0 and remain small in magnitude throughout the range of observed hourly cough counts. This motivates the use of two more CoughMonitor accuracy metrics:

- the *bias*, defined to be the average of the percentage errors of the CoughMonitor's hourly cough counts, and
- the *margin of error (MOE)* of the bias, defined to equal twice the standard deviation of the percentage errors of the CoughMonitor's hourly cough counts.

The bias and the MOE are summary statistics for Bland-Altman analyses; the Bland-Altman *limits of agreement* are the endpoints of the corresponding interval estimates, i.e., bias $\pm$ MOE. These metrics were calculated for the entire cohort and separately for each individual participant, and Bland-Altman plots of the hour-by-hour percentage errors against the ground truth hourly cough counts show how these differences depend on cough frequency.

2.4 Event-to-event analysis

While we maintain that the most clinically relevant metric of accuracy is correlation of hourly cough rates, we accept that hourly cough counts risk being overly reductive and that a high correlation between paired hourly cough counts could obscure inaccuracy at detecting individual events. To assess accuracy at the level of individual events, note that each individual CoughMonitor detection (cough or cough-second) is either a *true positive* or a *false positive*, by comparison with the ground truth annotations. We have then used the usual formulas to calculate the CoughMonitor's

- *sensitivity*, the number of true positives divided by the total number of ground truth events, and
- *hourly false positive rate (FPR)*, the number of false positives divided by the total number of hours of monitoring,

for both coughs and cough-seconds. We have also calculated the CoughMonitor's

- *specificity*, the number of false positives divided by the total number of possible false positives,

for both coughs and cough-seconds. When computing specificity, the denominator depends on whether coughs or cough-seconds are being counted; in the case of cough-seconds, there is 1 possible false-positive for every second without cough; in the case of coughs, the number of "negative" events is less clear; since coughs last on average approximately one-third of a second, we estimate that there are three possible false positives for every second without cough.

2.5 Primary and Secondary Objectives

Our primary objective is to assess the overall performance of HCMS when used by individuals as they go about their activities of daily living. Thus we first present this overall analysis of aggregate results.

We had three secondary objectives: 1) assessing performance during daytime and bedtime, 2) assessing performance for individuals, and 3) assessing performance as a function of cough rates.

Since the bedroom is an acoustically unique environment and the biological determinants and nature of coughing may be different during sleep, separate analyses were carried out for patients' daytime and nighttime hourly cough counts.

Given that individuals may have oddly sounding coughs or reside in an acoustically challenging environment, there may be person-to-person variability in the performance of HCMS. Thus, we performed individual-specific accuracy analyses.

Antitussive trials using intermittent short-term cough

monitoring have yielded perplexing results in those with cough rates below versus above 25 coughs per hour. Thus, we will assess the dependence of HCMS's performance on subjects' cough rates.

2.6 IRB

The protocol for this study was IRB-approved at both sites and participants gave their informed consent to be monitored, to adhere to study procedures, and to have their audio reviewed.

3 Results

3.1 Primary Objective: Aggregate results

3.1.1 Annotations and total counts

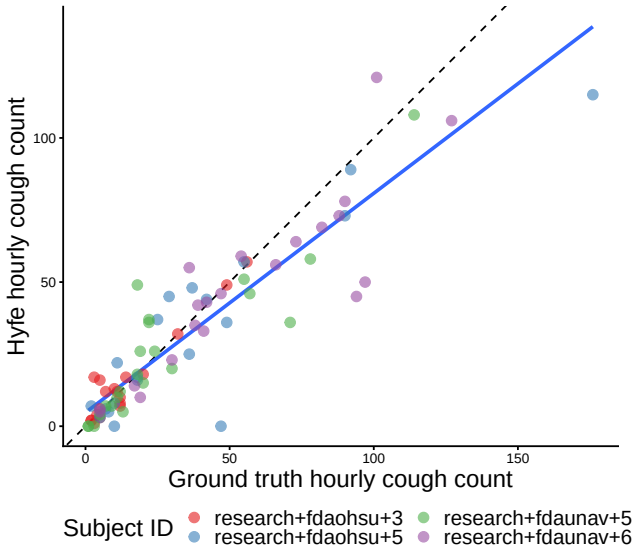
A total of 92.4 patient-hours were monitored and (double/triple) annotated. This was slightly less than the expected 96 patient-hours due to (a) patients returning to the clinic to return the device slightly before the 24-hour period was over and (b) one patient turning off his/her device at home. At the time of analysis, not all discrepancies had been resolved by the MD; in the case of non-resolved discrepancies, the most recent annotations for the disputed segments were used as "ground truth."

The following table shows the total number of coughs captured by the human annotators and by the CoughMonitor:

Ground truth	CoughMonitor
2832	2548

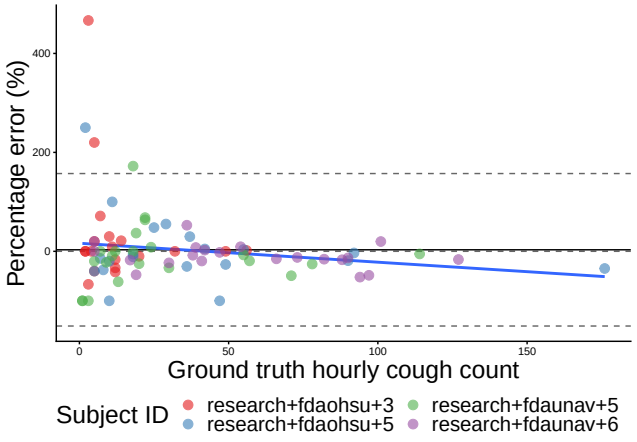
3.1.2 Linear analysis of coughs

The plot below shows the agreement between ground truth and CoughMonitor hourly cough counts. The Pearson correlation is 0.914 and the slope of the line of best fit (shown in blue) is 0.761. Each dot on the scatterplot is one person-hour.



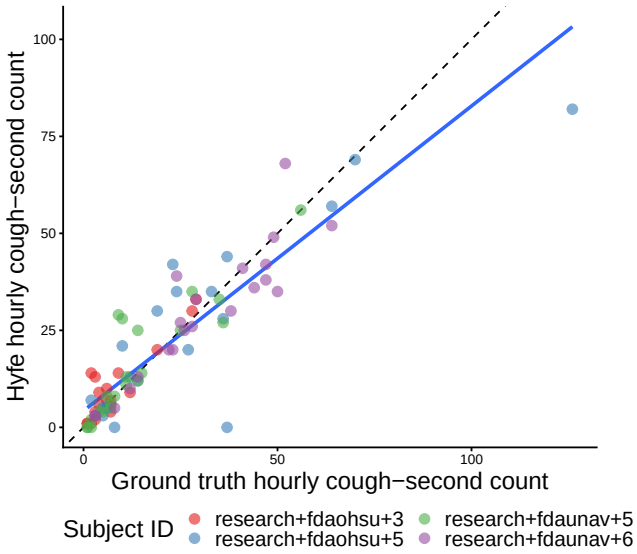
3.1.3 Bland-Altman analysis of coughs

Below is the Bland-Altman plot of percentage errors versus ground truth hourly cough counts. The bias is 3.116% (solid black line) and the slope of the line of best fit (shown in blue) is -0.383.



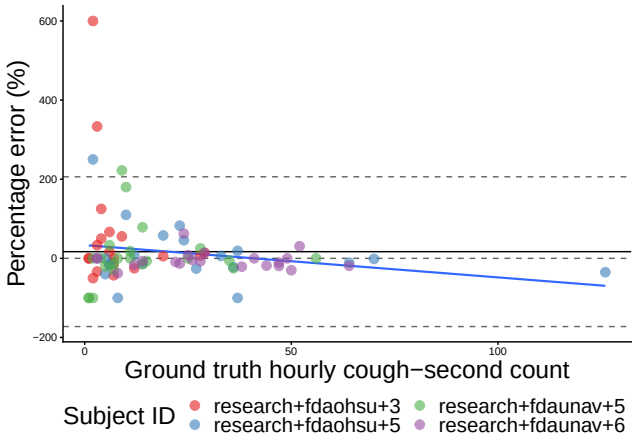
3.1.4 Linear analysis of cough-seconds

The plot below shows the agreement between ground truth and CoughMonitor hourly cough-second counts. The Pearson correlation is 0.899 and the slope of the line of best fit (shown in blue) is 0.786. Each dot on the scatterplot is one person-hour.



3.1.5 Bland-Altman analysis of cough-seconds

Below is the Bland-Altman plot of percentage errors versus ground truth hourly cough-second counts. The bias is 16.872% (solid black line) and the slope of the line of best fit (shown in blue) is -0.817.



3.1.6 Event-to-event analysis of coughs

The table below shows event-to-event accuracy for coughs.

Statistic	Value
True positives	1956.00
Predicted coughs	2548.00
False positives	592.00
True coughs	2832.00
False negatives	876.00
Sensitivity	69.07
Hours	92.38
False positives per hour	6.41
Specificity	99.94

3.1.7 Event-to-event analysis of cough-seconds

The table below shows event-to-event accuracy for cough-seconds.

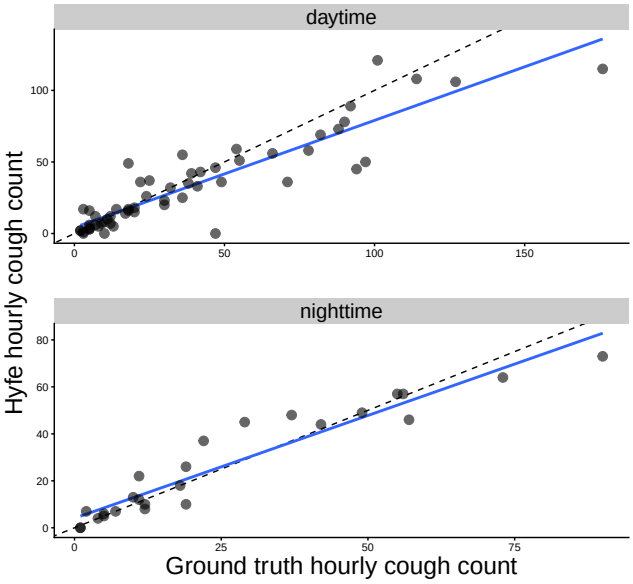
Statistic	Value
True positives	1281.00
Predicted coughs	1658.00
False positives	377.00
True coughs	1648.00
False negatives	367.00
Sensitivity	77.73
Hours	92.38
False positives per hour	4.08
Specificity	99.89

3.2 Secondary Objective: Daytime vs nighttime

3.2.1 Linear analysis of coughs, daytime vs nighttime

The following summarize and compare the daytime and nighttime agreement between ground truth and CoughMonitor hourly cough counts.

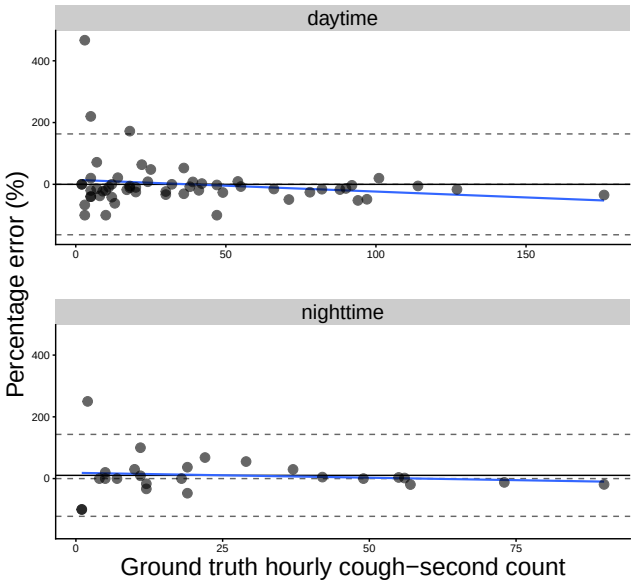
Period	Correlation	Intercept	Slope
nighttime	0.952	4.071	0.875
daytime	0.907	4.205	0.748



3.2.2 Bland-Altman analysis of coughs, daytime vs nighttime

Below are the Bland-Altman statistics and plots of percentage errors versus ground truth hourly cough counts, separating between day and night times.

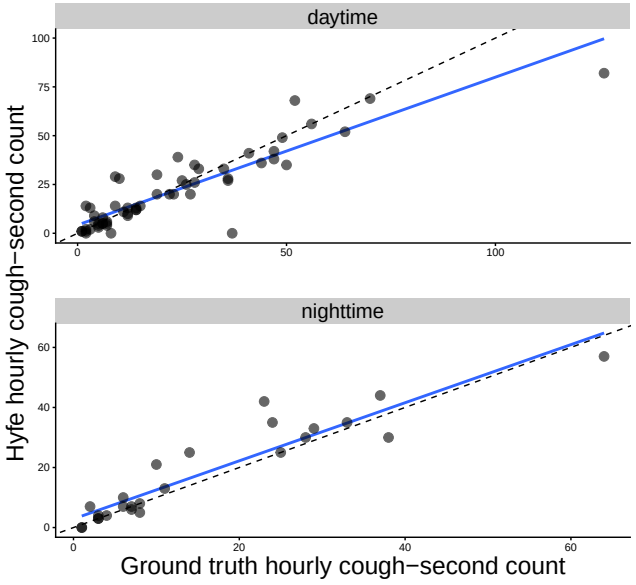
Period	Bias (%)	Bias MOE (%)
nighttime	10.453	132.689
daytime	-0.102	163.293



3.2.3 Linear analysis of cough-seconds, daytime vs nighttime

The following summarize and compare the daytime and nighttime agreement between ground truth and CoughMonitor hourly cough-second counts.

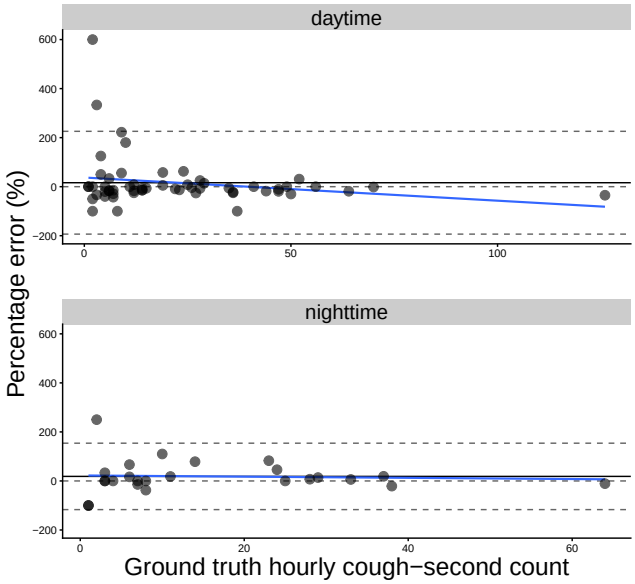
Period	Correlation	Intercept	Slope
nighttime	0.933	2.871	0.968
daytime	0.896	4.251	0.757



3.2.4 Bland-Altman analysis of cough-seconds, daytime vs nighttime

Below are the Bland-Altman statistics and plots of percentage errors versus ground truth hourly cough-second counts, separating between day and night times.

Period	Bias (%)	Bias MOE (%)
nighttime	18.560	135.437
daytime	16.132	209.663



3.2.5 Event-to-event analysis of coughs by individual, daytime vs nighttime

The tables below show event-specific results for each study participant by time of day, using coughs (not cough-seconds).

Table 1: Daytime

Subject ID	TP	Hyfe coughs	FP	True coughs	FN	Sensitivity	FP per hour	Specificity
research+fdaohsu+3	79	133	54	110	31	71.818	2.626	99.927
research+fdaohsu+5	312	361	49	501	189	62.275	2.045	99.943
research+fdaunav+5	315	437	122	485	170	64.948	5.087	99.858
research+fdaunav+6	759	948	189	1089	330	69.697	7.919	99.777

Table 2: Nighttime

Subject ID	TP	Hyfe coughs	FP	True coughs	FN	Sensitivity	FP per hour	Specificity
research+fdaohsu+3	105	154	49	154	49	68.182	2.382	99.934
research+fdaohsu+5	212	296	84	266	54	79.699	3.505	99.902
research+fdaunav+5	99	134	35	125	26	79.200	1.459	99.959
research+fdaunav+6	75	85	10	102	27	73.529	0.419	99.988

3.2.6 Event-to-event analysis of cough-seconds by individual, daytime vs nighttime

The table below shows event-specific results for each study participant by time of day, using cough-seconds (not coughs).

Table 3: Daytime

Subject ID	TP	Hyfe coughs	FP	True coughs	FN	Sensitivity	FP per hour	Specificity
research+fdaohsu+3	52	94	42	67	15	77.612	2.042	99.981
research+fdaohsu+5	236	272	36	368	132	64.130	1.502	99.986
research+fdaunav+5	187	268	81	240	53	77.917	3.377	99.969
research+fdaunav+6	473	572	99	583	110	81.132	4.148	99.962

Table 4: Nighttime

Subject ID	TP	Hyfe coughs	FP	True coughs	FN	Sensitivity	FP per hour	Specificity
research+fdaohsu+3	67	97	30	84	17	79.762	1.459	99.986
research+fdaohsu+5	167	240	73	193	26	86.528	3.046	99.972
research+fdaunav+5	60	75	15	64	4	93.750	0.625	99.994
research+fdaunav+6	39	40	1	49	10	79.592	0.042	100.000

3.3 Secondary objective: Individual results

Disease diagnoses The table below shows the disease diagnosis for each study participant.

hyfe_id	Diagnosis
research+fdaohsu+3	Bronchiectasis/NTM
research+fdaohsu+5	Bronchiectasis
research+fdaunav+5	Chronic cough
research+fdaunav+6	Post-infectious cough

Cough counts The table below shows total cough counts for each study participant.

Subject ID	Ground truth	CoughMonitor
research+fdaohsu+3	264	286
research+fdaohsu+5	767	653
research+fdaunav+5	610	571
research+fdaunav+6	1191	1033

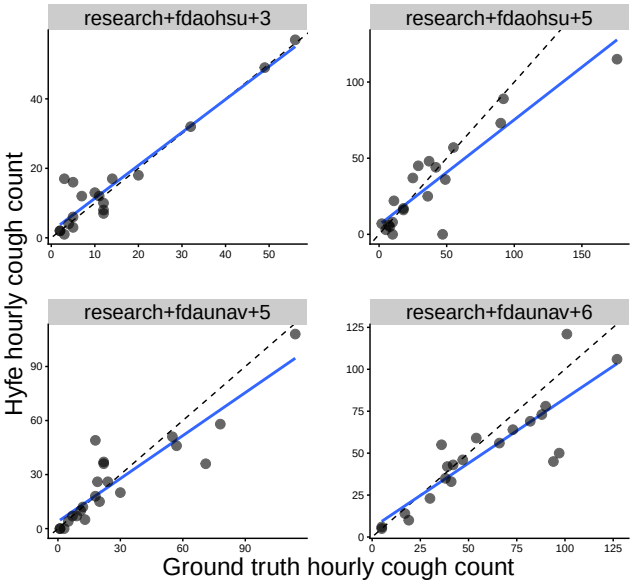
Cough-second counts The table below shows total cough-second counts for each study participant.

Subject ID	Ground truth	CoughMonitor
research+fdaohsu+3	153	191
research+fdaohsu+5	564	513
research+fdaunav+5	307	345
research+fdaunav+6	649	615

3.3.1 Linear analysis of coughs by individual

The following summarize and compare the agreement between ground truth and CoughMonitor hourly cough counts for each individual.

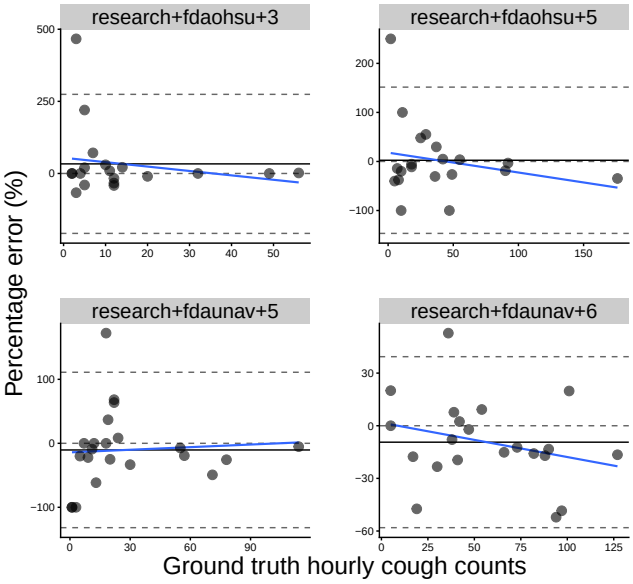
Subject ID	Correlation	Intercept	Slope
research+fdaohsu+3	0.954	1.860	0.949
research+fdaohsu+5	0.903	6.157	0.691
research+fdaunav+5	0.901	3.896	0.796
research+fdaunav+6	0.874	5.481	0.771



3.3.2 Bland-Altman analysis of coughs by individual

Below are the Bland-Altman statistics and plots of percentage errors versus ground truth hourly cough counts for each participant.

Subject ID	Bias (%)	Bias MOE (%)
research+fdaohsu+3	33.267	241.001
research+fdaohsu+5	2.448	149.116
research+fdaunav+5	-10.397	121.457
research+fdaunav+6	-9.371	48.725

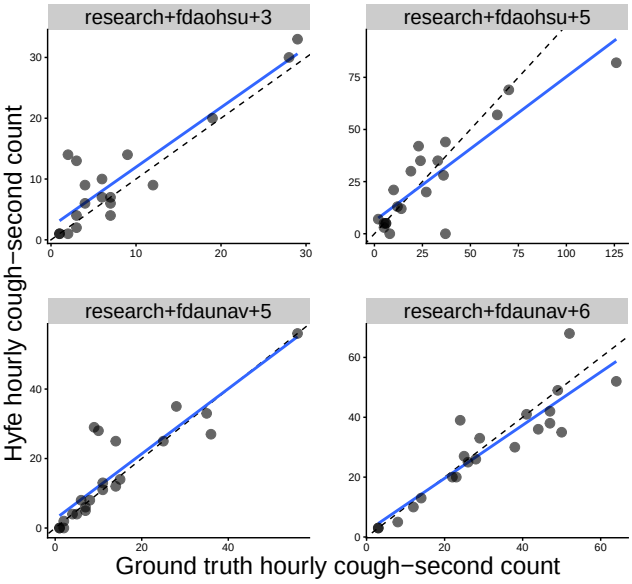


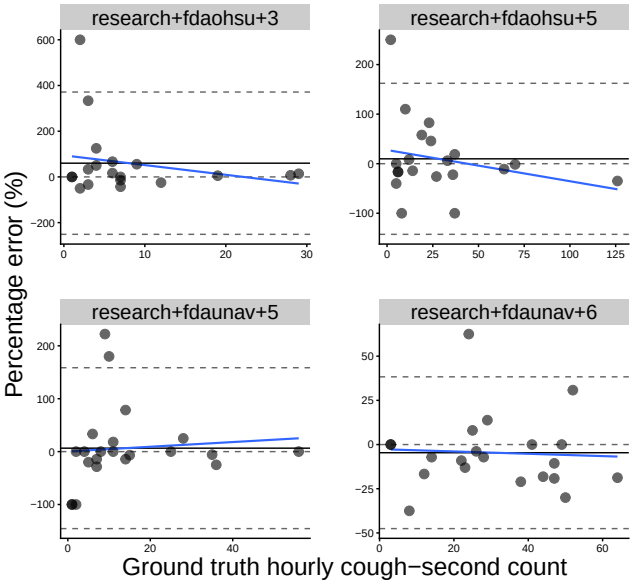
3.3.3 Linear analysis of cough-seconds by individual

The following summarize and compare the agreement between ground truth and CoughMonitor hourly

cough-second counts for each individual.

Subject ID	Correlation	Intercept	Slope
research+fdaohsu+3	0.901	2.153	0.981
research+fdaohsu+5	0.870	6.147	0.692
research+fdaunav+5	0.892	2.602	0.937
research+fdaunav+6	0.909	1.850	0.888





3.3.4 Bland-Altman analysis of cough-seconds by individual

Below are the Bland-Altman statistics and plots of percentage errors versus ground truth hourly cough-second counts for each participant.

Subject ID	Bias (%)	Bias MOE (%)
research+fdaohsu+3	60.067	311.323
research+fdaohsu+5	9.830	152.367
research+fdaunav+5	6.490	152.079
research+fdaunav+6	-4.626	42.934

3.3.5 Event-to-event analysis of coughs by individual

The table below shows event-specific results for each study participant, using coughs (not cough-seconds).

Subject ID	TP	Hyfe coughs	FP	True coughs	FN	Sensitivity	FP per hour	Specificity
research+fdaohsu+3	184	287	103	264	80	69.697	5.008	99.860
research+fdaohsu+5	524	657	133	767	243	68.318	5.549	99.844
research+fdaunav+5	414	571	157	610	196	67.869	6.546	99.817
research+fdaunav+6	834	1033	199	1191	357	70.025	8.338	99.765

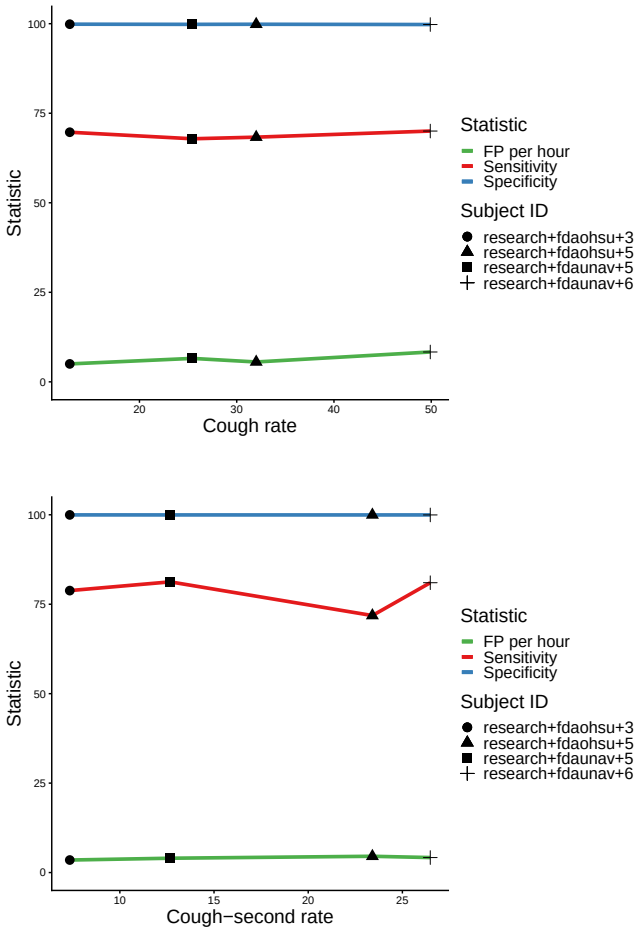
3.3.6 Event-to-event analysis of cough-seconds by individual

The table below shows event-specific results for each study participant, using cough-seconds (not coughs).

Subject ID	TP	Hyfe coughs	FP	True coughs	FN	Sensitivity	FP per hour	Specificity
research+fdaohsu+3	119	191	72	151	32	78.808	3.501	99.968
research+fdaohsu+5	403	512	109	561	158	71.836	4.548	99.958
research+fdaunav+5	247	343	96	304	57	81.250	4.003	99.963
research+fdaunav+6	512	612	100	632	120	81.013	4.190	99.961

3.4 Secondary objective: Performance by Cough Rate

The following plots show that performance does not depend on cough frequency; performance is similar for all pilot subjects, regardless of cough rate or unit of analysis (coughs or cough-seconds).



4 Discussion

Here we report results of a pilot evaluation of the HCMS, detailing the manner in which we established the ground truth and compared this to the results from a Version 0.0.1 of the HCMS. This White Paper serves as a preview of the performance we anticipate with Ver 1.0.0 and the format in which those results will be published and provided to the FDA.

In brief, we recorded a day in the life of four research subjects with problematic coughs while monitoring them with HCMS (version 0.0.1). Each recording was listened to by 2 trained cough annotators. Coughs were time stamped using proprietary annotation software according to detailed standard operating procedures. Any discrepancies between these annotators was adjudicated by an individual with special experience in cough annotation. The intra and inter observer consistency of this annotation process has been previously demonstrated to be very high (unpublished data available upon request). The resulting time stamps obtained from trained human annotators were considered the ground truth for coughing.

Coughs were analyzed based on two complementary units of analysis - individual coughs and cough-seconds. A cough-second is a second during which at least one individual cough occurs. While there are technical reasons that favor one or the other, both are interchangeable from the perspective of the end users and can ultimately be described as "hourly cough counts".

The person-hour by person-hour performance of the HCMS was quantified by comparing hourly cough rates (presented with linear and Bland Altman statistics) and event by event by comparing individual coughs or cough seconds (presented as sensitivity, specificity and false positive rates). Although we employed and reported all methods and we consider them to be complementary, for communication purposes only a subset of this analysis will be cited in making a performance claim.

This panel of analysis was used to assess the primary endpoint which was the overall performance of the HCMS when used by problematic coughers under common living conditions. A total of 2832 coughs were recorded over the 92.4 hours monitored. Linear analysis for both cough and cough seconds were high with correlations of 0.91 and 0.90 and slopes of 0.76 and 0.79 respectively. Similarly, event by event analysis of cough and cough seconds showed sensitivity of 69% and 78% and false positivity rates of 6.4 and 4.0 coughs per hour.

This panel was used for the secondary objectives which explore in further detail subsets of the data. Specifically:

There was a difference in performance between daytime and bedtime, with correlations of 0.91 and 0.95 and slopes of 0.75 and 0.88, respectively. With the event to event analysis the improvement at night time was even more pronounced in two of the four individuals with cough sensitivity improved as much as 17% and false positivity rate decreased as much as 7.5 coughs per hour. This was also true with cough seconds where sensitivity increased as much as 22% and false positivity decreased as much as 2.7 coughs per hour.

Coughing differed significantly between individuals, ranging from 264 to 1191 coughs over the day-long recording period. However the performance was rather consistent, ranging between correlation of .87 to .95 and slopes of .77 to .95. When compared event by event the sensitivity ranged from 67% to 70% and the false positivity rate from 6.5 to 8.3 coughs per hour.

There was no significant difference in the accuracy of HCMS as a function of cough or cough second rates. This suggests that HCMS can be used for patients and cohorts irrespective of their couth rates.

5 Conclusion

In conclusion, we have established a rigorous methodology for evaluating AI enabled cough monitoring, reporting their results and establishing claims as to its performance. Unsurprisingly, the HCMS performs better in the quiet background of nightime. By piloting this methodology with a prior version of the HCMS (0.0.1) we have gained confidence that we will achieve clinically meaningful performance with the current Version 1.0.0. Taken together, this is encouraging data suggesting that scientifically driven development and evaluation may soon bring continuous passive cough monitoring into service to improve healthcare delivery.