RESEARCH ARTICLE

REVISED Performance evaluation of the smartphone-based AI cough monitoring app - Hyfe Cough Tracker against solicited respiratory sounds [version 2; peer review: 1 approved with reservations, 1 not approved]

Mindaugas Galvosas[1*], Juan C. Gabaldón-Figueira[2*], Eric M. Keen[1], Virginia Orrillo[3], Isabel Blavia[3], Juliane Chaccour[2], Peter M. Small[1,4], Gerard Giménez[1], Matthew Rudd[1,9], Simon Grandjean Lapierre[5,6*], Carlos Chaccour[2,7,8*]

[1]Research and Development Department, Hyfe Inc., Wilmington, Delaware, USA
[2]Department of Microbiology and Infectious Diseases, Clinica Universidad de Navarra, Pamplona, Spain
[3]School of Pharmacy and Nutrition, University of Navarra, Pamplona, Spain
[4]Department of Global Health, University of Washington, Seattle, Washington, USA
[5]Immunopathology Axis, Research Center, University of Montreal Hospital Center, Montreal, Canada
[6]Department of Microbiology, Infectious Diseases and Immunology, University of Montreal, Montreal, Canada
[7]ISGlobal, Hospital Clinic, University of Barcelona, Barcelona, Spain
[8]Centro de Investigación Biomédica en Red de Enfermedades Infecciosas, Madrid, Spain
[9]Department of Mathematics and Computer Science, Sewanee The University of the South, Sewanee, Tennessee, USA

[*] Equal contributors

v2 **First published:** 01 Jul 2022, **11**:730
https://doi.org/10.12688/f1000research.122597.1
Latest published: 09 Jun 2023, **11**:730
https://doi.org/10.12688/f1000research.122597.2

Abstract

Background: Emerging technologies to remotely monitor patients' cough show promise for various clinical applications. Currently available cough detection systems all represent a trade-off between convenience and performance. The accuracy of such technologies is highly contingent on the clinical settings in which they are intended to be used. Moreover, establishing gold standards to measure this accuracy is challenging.

Objectives: We present the first performance evaluation study of the Hyfe Cough Tracker app, a passive cough monitoring smartphone application. We evaluate performance for cough detection using continuous audio recordings obtained within a controlled environment and cough counting by trained individuals as the gold standard. We propose standard procedures to use multi-observer cough sound annotation from continuous audio recordings as the gold standard for evaluating automated cough detection devices.

Methods: This study was embedded in a larger digital acoustic

Open Peer Review

Approval Status ? ✗

	1	2
version 2 (revision) 09 Jun 2023		
version 1 01 Jul 2022	? view	✗ view

1. **Vasileios Papapanagiotou**, Aristotle University of Thessaloniki, Thessaloniki, Greece

2. **Terence E. Taylor**, Vitalograph Ireland Ltd., Ennis, Ireland

surveillance study (clinicaltrial.gov NCT04762693). Forty-nine participants were included and instructed to produce a diverse series of solicited sounds in 10-minute sessions. Simultaneously, continuous audio recording was performed using a MP3 recorder and two smartphones running Hyfe Cough Tracker app monitored and identified cough events. All continuous audio recordings were independently labeled by three medically-trained researchers.
Results: Hyfe Cough Tracker app showed sensitivity of 91% and specificity of 98% with a very high correlation between the cough rate measured by Hyfe and that of human annotators (Pearson correlation of 0.968). A standardized approach to establish an acoustic gold standard for identifying cough sounds with multiple observers is presented.
Conclusion: This is the first performance evaluation of a new smartphone-based cough monitoring system. Hyfe Cough Tracker can detect, record and count coughs from solicited cough-like explosive sounds in controlled acoustic environments with very high accuracy. Additional steps are required to validate the system in clinical and community settings.

Keywords
cough, artificial intelligence, cough monitoring, cough counting, hyfe, hyfe cough tracker

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Mindaugas Galvosas (mindaugas.g@hyfe.ai)

Author roles: Galvosas M: Data Curation, Formal Analysis, Project Administration, Resources, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Gabaldón-Figueira JC**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Keen EM**: Formal Analysis, Methodology, Visualization, Writing – Review & Editing; **Orrillo V**: Investigation, Writing – Review & Editing; **Blavia I**: Investigation, Writing – Review & Editing; **Chaccour J**: Investigation, Writing – Review & Editing; **Small PM**: Writing – Review & Editing; **Giménez G**: Formal Analysis, Writing – Review & Editing; **Rudd M**: Formal Analysis, Visualization, Writing – Review & Editing; **Grandjean Lapierre S**: Funding Acquisition, Supervision, Writing – Review & Editing; **Chaccour C**: Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: MG, EK, GG, PM and MR are employees of Hyfe Inc. Hyfe had no role in the decision to submit this protocol for publication. CCH has received consultancy fees and owns equity from Hyfe Inc. No competing interests were disclosed for all other authors.

Grant information: This study was funded by the Patrick J. McGovern Foundation. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. ISGlobal acknowledges support from the Spanish Ministry of Science and Innovation through the "Centro de Excelencia Severo Ochoa 2019-2023" Program (CEX2018-000806-S), and support from the Generalitat de Catalunya through the CERCA Program.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Galvosas M, Gabaldón-Figueira JC, Keen EM *et al.* **Performance evaluation of the smartphone-based AI cough monitoring app - Hyfe Cough Tracker against solicited respiratory sounds [version 2; peer review: 1 approved with reservations, 1 not approved]** F1000Research 2023, **11**:730 https://doi.org/10.12688/f1000research.122597.2

First published: 01 Jul 2022, **11**:730 https://doi.org/10.12688/f1000research.122597.1

REVISED **Amendments from Version 1**

The updated version includes clarifications that this performance evaluation study was done in controlled environment, and it was the sole purpose of this evaluation, informing further decisions leading to a separate trial and publication in regard to validating the Hyfe Cough Monitoring system in real world environments. We appreciate relevant reviewers' comments and provided point by point answers, additionally implementing edits in the parts of "Abstract - objectives", "Methods - Automated cough detection system", "Study design", "Results" and "Discussion". We are adding two new tables on sensitivity, specificity statistics and linear analysis model parameter estimates. Additionally, we are replacing Figure 2 with an updated quality image and adding additional figures for Bland-Altman plots and linear analysis plots for Human annotated coughs and Hyfe detected coughs. Matthew Rudd is added as a new co-author, since his cough-data science contribution was required in making an additional analysis, visualisations and advice was taken on responses to reviewers and edits to data-related parts of the manuscript. Github code was updated to include the additional analysis.

Any further responses from the reviewers can be found at the end of the article

Introduction

Cough is consistently ranked as one of the most common reasons for seeking medical attention.[1,2] Acute cough frequently indicates new-onset and potentially contagious respiratory infection,[3] while chronic cough can be an important cause of discomfort and disability affecting quality of life.[4,5] In current medical practice, objective cough assessment can only occur during face to face interaction with the patient in the context of in- or outpatient visits, effectively making the symptom invisible to the health care provider outside the medical settings. To assess cough in ambulatory settings, health care providers rely on questionnaires and patient-reported outcomes, which are subject to patients' self-perception, cough tolerance and recall bias.[6,7] While different systems for automated cough detection have been developed in the last decade,[8,9] they depend on wearable microphones, or spirometers,[10] and their adoption is limited by cost, portability and privacy concerns given the need for continuous sound recording. Recent advances in artificial intelligence (AI) allow the monitoring of cough in a non-obtrusive way using smartphones or other wearable digital devices.[6,11–14] Unobtrusive and privacy preserving passive cough monitors could revolutionize clinical practice and research in the field of respiratory diseases.

Longitudinal monitoring of cough is particularly attractive for the evaluation of disease progression, or treatment response, as well as in clinical trials where trends in cough rates is an outcome of interest. Longitudinal cough monitoring also opens the door to population-wide capture of cough-signals as a surrogate marker of respiratory diseases epidemiology.[14]

Evaluating cough and its patterns with limited recording periods (e.g., 24 h) can be misleading, in particular, if only small changes in cough frequency are captured over the limited 24 h recording and in cases that have high variance of cough counts.[12] However, the nature and volume of data generated with protracted monitoring raises new challenges in technology validation. A central challenge in this work is establishing a gold standard against which automated devices' performance can be evaluated.

In this study, we present the accuracy of Hyfe Cough Tracker app (henceforth referred to as Hyfe), a smartphone-based automated cough monitor that uses a convolutional neural network (CNN) to differentiate coughs from other explosive sounds.[13] In this "in-vitro" performance evaluation, we use solicited sounds in a controlled acoustic environment as the first step towards clinical validation. We also propose a standard operating procedure (SOP) to appropriately label cough sounds from continuous audio recording.

Methods

Automated cough detection system

Hyfe is a software application for patient use, freely available for use on Android and iOS smartphones. It continuously monitors ambient sound and employs a two-step process to (1) detect explosive cough-like sounds, record a 0.5 second sound snippet which is sent to a cloud server where (2) a CNN assigns a cough prediction score (0 to 1) to each sound. Hyfe's CNN model, at the time of this analysis, was trained on more than 200M real-world cough and cough-like samples, collected from multiple countries and multiple mic-enabled devices. For this study, a minimal score threshold of 0.85 was used for classifying a peak sound as a cough. Within this study, Hyfe (Version acl 1.24.4) was installed on smartphones (Motorola G30, Motorola, Inc, Chicago, IL, USA) running Android operating system version 11 (Google, LLC, Mountainview, CA, USA).

To assess the accuracy of Hyfe, continuous recording using a MP3 recorder (Sony ICD-PX470, Sony, Tokyo, Japan) and manual labeling of cough by medically trained listeners was used as the gold standard.

Study design

This performance evaluation study was conducted at the University of Navarra, Spain, between September to November 2021 and was nested in a larger cohort study (Clinicaltrials.org NCT04762693).[12] Both the main and the nested study received approval by the Medical Research Ethics Committee of the chartered community of Navarra (PI_2020/107). Students and staff from the university of Navarra were invited to participate via email. All participants were aged 18 or older and signed informed consent. Baseline respiratory symptoms were not considered for inclusion. Participants were asked to produce a series of solicited sounds by reading a provided script, while being recorded with an MP3 recorder and monitored by Hyfe on two identical smartphones. The phones and recorder were placed on a table at approximately 50 cm from the participants, with microphones oriented towards them.

A pre-generated computer script instructed participants to produce a series of 46 sounds, of which 18 were coughs, the rest consisted of solicited sneezes, throat clearings, spoken letters or words in the same 10 minutes. Participants were instructed to cough once every time they were prompted by the script to do so. In total for each participant, the script included instructions to cough 20 (18 as isolated coughs and 2 coughs in the literary text) times, sneeze 10 times, clear their throat 5 times and produce 15 sounds (explosive words, for example, "paella" and numbers as "93"). Some sounds were requested while reading out loud a literary text (in Spanish). Outside the reading, solicited sounds were separated from one another by at least five seconds of ambient silence. There were five different versions of the script, each one presenting a different sequence of instructions, and the version shown to each participant was randomly selected using a computer-generated sequence at the beginning of each session. Recording sessions occurred in a quiet room and lasted approximately 10 minutes. The sampling rate was 44.1 Hz and the files are 16-bit. The time at which individual sounds were produced was automatically recorded in every session. Sound intensity levels in the room were also monitored using a UNI-T mini sound level meter. The room was not acoustically insulated.

Three medically trained researchers listened to individual recording sessions using Audacity (Audacity team (2021). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.1.3).[15] Coughs were manually annotated using digital audio recordings and visual audio wave representation. It was previously shown that ambulatory cough counts from audio recordings have great agreement with patient video recordings, and that digital audio recordings could hence be considered as the gold standard in validating novel cough monitoring tools.[16,17] Each sound was labeled using a 4-tier system defined in the SOP, which was developed for cough annotation in continuous audio recordings. In brief, sounds were classified as 0 = definitely not a cough, 1 = disputable cough (i.e., someone could consider the sound as a cough), 2 = definite cough but distant/muffled/obstructed, 3 = definite cough. Labels were made using Audacity and exported as text files for analysis. Labellers were blinded to the classification made by Hyfe and other listeners but knew a participant's age and gender. Sounds labeled unanimously as a number 3 ("definite cough") by all the human listeners were considered true coughs.

Sample size

We estimated that at least 385 sounds would be required to observe a 90% sensitivity and 85% specificity, with a cough prevalence of 40% (39% of solicited sounds in the script were coughs), a precision of 5%, and a dropout rate of 10%.[18,19]

Data processing and analysis

Labels created by listeners (in Audacity) and Hyfe detected coughs were firstly manually synchronized to within two seconds (as this was within the silent time of five seconds between the solicited sounds in the automated script) of each other. Synchronization was then carried out for each phone and each session separately by identifying the time offset that would align Hyfe detections with the labels and adjusting the Hyfe detection timestamps accordingly. Offsets were estimated first using a subroutine in R that iteratively tests the offset-error produced by a wide variety of values, then manually reviewing and adjusting those automatic offsets as needed.

For the performance analysis, each recording session was divided into seconds. Seconds in which at least one explosive cough-like sound was labeled by a human listener (categories 1, 2, or 3) were pooled and defined as "cough-like-seconds". Individual labels, which were annotated by the listeners, occurring within one second of each other were treated as a single label, and included as a single "cough-second". Similarly, seconds in which only non-cough sounds occurred (category 0), were identified as "non-cough seconds".

Hyfe detections on each phone were also pooled into cough-seconds using a similar method: all detected explosive sounds occurring within a one-second period were treated as a single detection; if multiple explosive sounds occurred within a cough-second, the highest cough prediction score among all explosive phases was used as the prediction score for the cough-second.

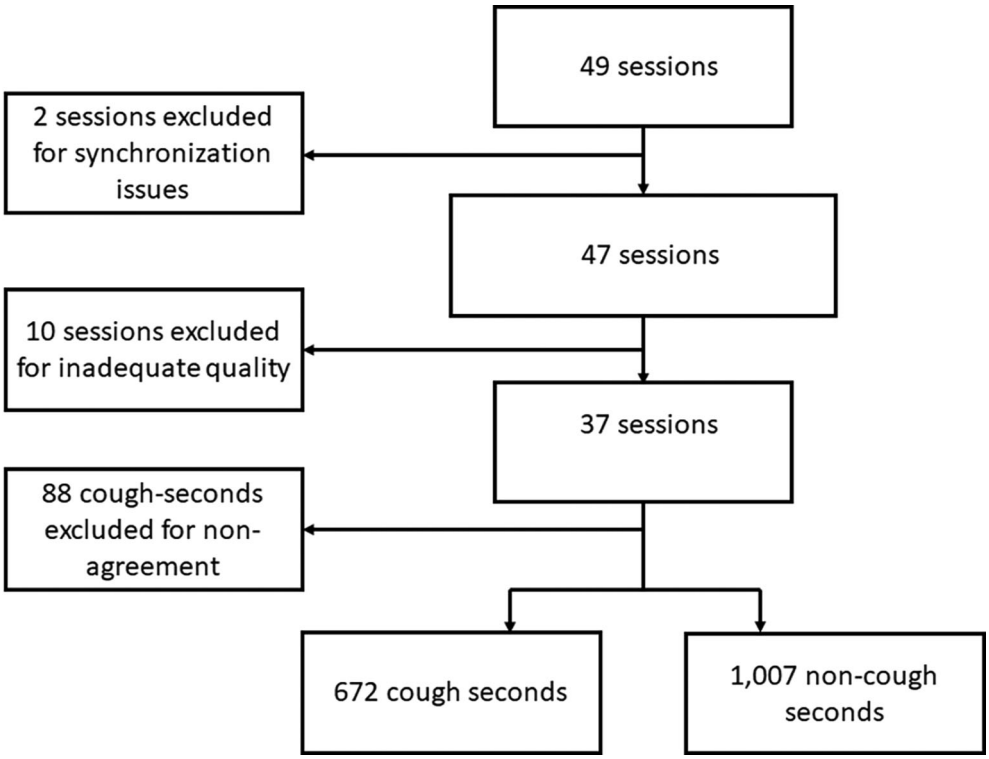

Figure 1. Study flow chart.

All recording seconds were considered as distinct analysis units. Seconds for which there was disagreement between the three human listeners were excluded from the final analysis. Similarly, 10-minute sessions in which fewer than 10 sounds were unanimously labeled as coughs by human listeners were considered of inadequate quality and excluded (Figure 1). True positives (TP) assessments were defined as those cough-seconds detected by Hyfe, and unanimously classified as category 3 by all human listeners. False positives (FP) were defined as seconds in which coughs did not actually occur, but were incorrectly detected by Hyfe. A pooled sensitivity and specificity value for each phone was obtained by aggregating the cough- and non-cough seconds labeled and detected by each phone throughout all sessions included in the analysis. The fraction of TP among cough-seconds was calculated (sensitivity), as well as the fraction of FP among all non-cough seconds, which was used to calculate the specificity, using the following formula: 1 - (FP/non-coughs) = Specificity.

Given inter-participant variation in ability to generate coughs and other sounds, the performance characteristics of Hyfe for each combination of phone and session were individually assessed and then used to calculate an average sensitivity and specificity in an exploratory sub-analysis.

All data processing and analysis was performed in R version 4.02 (R Core Team 2020) and the code used is available from GitHub and is archived with Zenodo.[23]

This analysis further informed the SOP used by Hyfe to annotate coughs and cough-like sounds (sneezes and throat clears), leading to the most recent version - the 6-tier SOP for cough labeling in continuous audio recordings, which now also instructs to label the complete duration of target sounds.

Because the utility of a cough monitor is not in noting individual coughs but rather in tracking cough rates, we further analyzed these results to look at the overall performance of Hyfe to the human annotated gold standard. We cut the entire observation period for all participants into one-minute segments, then compared the gold standard (the number of coughs during that minute per the human annotator) against the tool (the number of cough detections per Hyfe).

Results

In total, 49 recording sessions with individual participants of approximately 10 minutes each were carried out. Two sessions did not have enough labels or detections to allow adequate timestamp synchronization and were excluded.

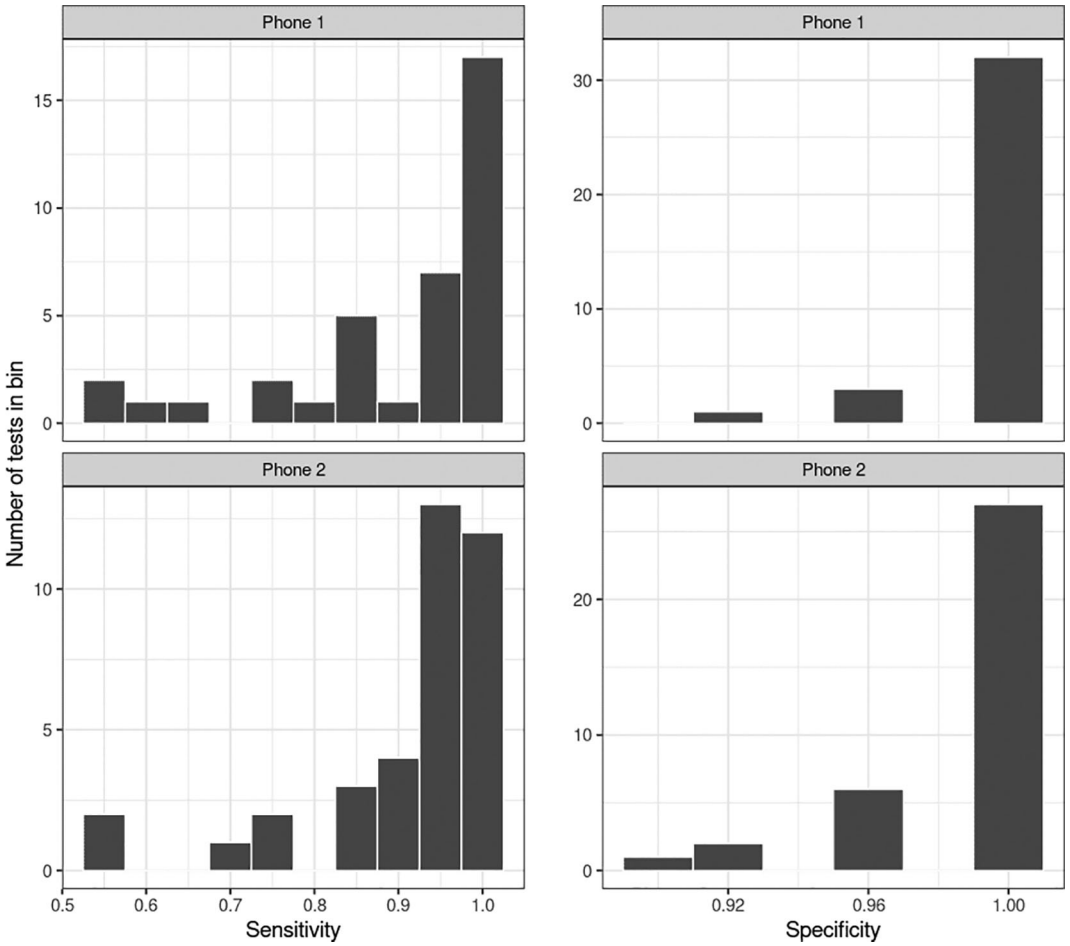

Figure 2. Performance of both phones through the 37 studied sessions. Sensitivity and specificity of Hyfe Cough Tracker assessed using solicited coughs.

Ten sessions did not have at least 10 sounds unanimously labeled as coughs and were also excluded, leaving 37 sessions, with 672 unanimously-labeled cough-seconds, and 1,007 non-cough seconds for the final performance evaluation (Figure 1).

The performance of Hyfe using both phones was similar in the pooled analysis, and is presented in Figure 2. Summary statistics of separate tests on sensitivity and specificity are presented in Table 1, showing the median 0.944 sensitivity and median 1.000 specificity for both phones. In a pooled analysis, Phone 1 yielded a sensitivity of 91.5% (95% CI: 89.2%-93.5%) and a specificity of 99.3% (95%CI: 98.6%-99.7%, Table 2), while phone 2 yielded a sensitivity of 92.55% (95% CI: 90.3%-94.4%) and a specificity of 98.7% (95% CI: 97.8%-99.3%, Table 2). The performance of both phones in individual sessions was also evaluated - the average sensitivity of the system in both phones and through the 37 sessions was 90.8% (SD = 11.6%). Specificity was high in both phones (range 93%-100% for phone 1, and 89%-100% for phone 2), with the mean specificity being 99.1% (SD = 1.9%). Sound levels in the room during the study were never above 110dB.

In three recording sessions, Hyfe had a sensitivity around 55%: sessions 2, 17 and 38 (Figure 3). These sessions met the quality criteria of more than 10 sounds unanimously classified as coughs. Potential explanations for this performance include the acoustic characteristics of the solicited coughs from these particular participants and the level of background noise. Coughs in session 2 and 17 had uncommon acoustic characteristics, such as biphasic decibel peaks, and different spectrographic features. Session 38 had significantly more background noise than the others. Sensitivity for the Session 20 was not evaluated because this was a patient with refractory chronic cough that generated hundreds of out-of-script, making timestamping impossible. We found the Pearson correlation of Hyfe to the gold standard to be 0.968 (Figure 4) with an intercept of -3.535 and slope of 1.214 for Phone 1, and intercept of -3.248 and slope of 1.213 for Phone 2 (Table 3).

Table 1. **Summary statistics on sensitivity and specificity for both phones used in individual sessions.**

	Minimum	First quartile	Median	Third quartile	Maximum
Sensitivity					
Phone 1	0.556	0.867	0.944	1.000	1.000
Phone 2	0.545	0.889	0.944	1.000	1.000
Specificity					
Phone 1	0.929	1.000	1.000	1.000	1.000
Phone 2	0.893	0.991	1.000	1.000	1.000

Table 2. **Comparative performance of both phones used.**

		Human labels					
		Phone 1			Phone 2		
		Cough seconds	Non-cough seconds	Total	Cough seconds	Non-cough seconds	Total
Hyfe's classification	Cough seconds	615	7	622	622	13	635
	Non-cough seconds	57	1000	1057	50	994	1044
	Total	672	1007	1679	672	1007	1679

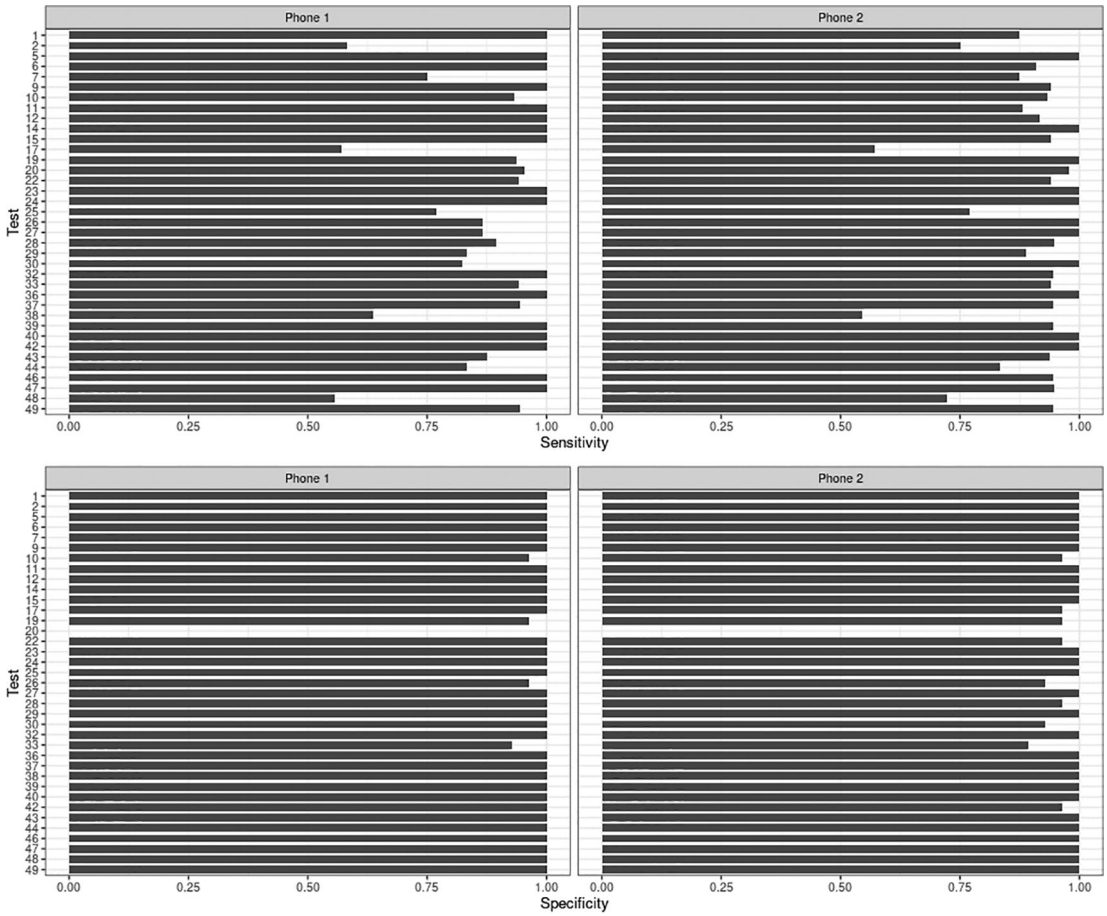

Figure 3. **Performance of the Hyfe Cough Tracker in individual recording sessions.**

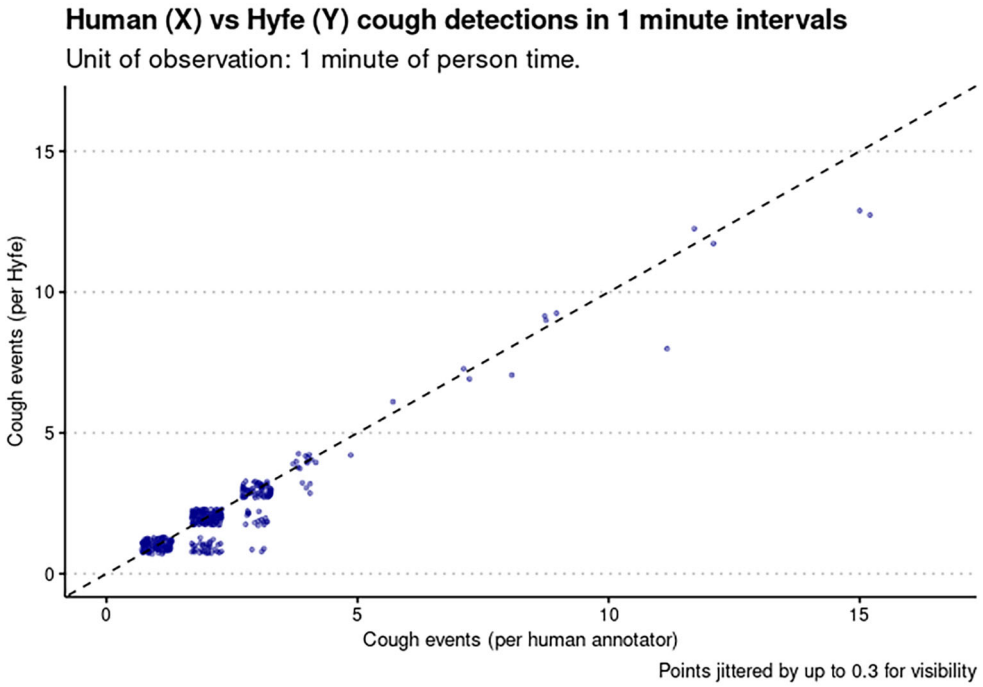

Figure 4. Correlation between the gold standard (human annotator) on the x-axis and the monitor (Hyfe) on the y-axis. Points are intentionally jittered by up to 0.3 values so as to provide more visibility on high density areas. The diagonal line (slope = 1, intercept = 0) represents where each point would fall in the hypothetical case of a perfect monitor.

Table 3. Linear analysis on model parameter estimates for both phones used.

	Linear model parameter estimates		
	Pearson correlation	**Intercept**	**Slope**
Phone 1	0.986	-3.535	1.214
Phone 2	0.989	-3.248	1.213

The linear analysis (Figure 5) and Bland-Altman plot based on percentage error (Figure 6) for the agreement of human annotated coughs and Hyfe cough detections are also presented.

Limitations

The major limitation was that this study of performance evaluation was done in a laboratory "in-vitro" environment, not community or a clinical setting. During this study, phone microphones were oriented towards and phones were placed at 50 cm from the participant, however, these settings would vary in real life clinical scenarios with coughing patients which could have longer distances and obstructing objects in between.

Discussion

The ability to unobtrusively monitor cough has the potential to greatly improve patient care, public health and drug development. The uptake of cough monitoring technologies will be determined by their usability, their clinical performance and the increasing evidence that they can provide actionable information for clinical decision making. Hyfe has advantages over existing cough monitors as it can run in the background of a smartphone and passively monitor coughs for longer than 24h of recordings. Rather than using special equipment and limited time windows for continuous cough monitoring, the use of this novel system improves the efficiency of monitoring and reduces the monitoring costs.

There are many ways to assess cough detectors accuracy. The intrinsic, or analytical, performance of AI-based cough monitors directly results from their algorithm's sensitivity and specificity for labeling recorded sounds. However, those same monitoring technologies may perform differently when deployed in various clinical settings where the acquisition

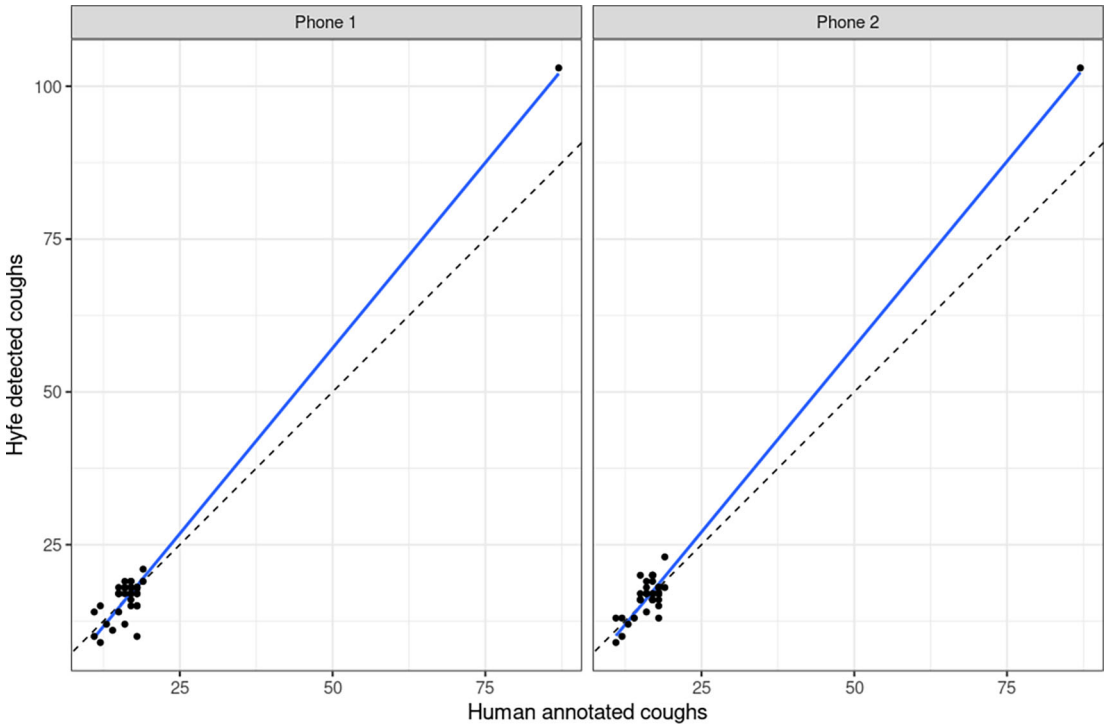

Figure 5. **Linear analysis plot of Human annotated coughs and Hyfe detected coughs.**

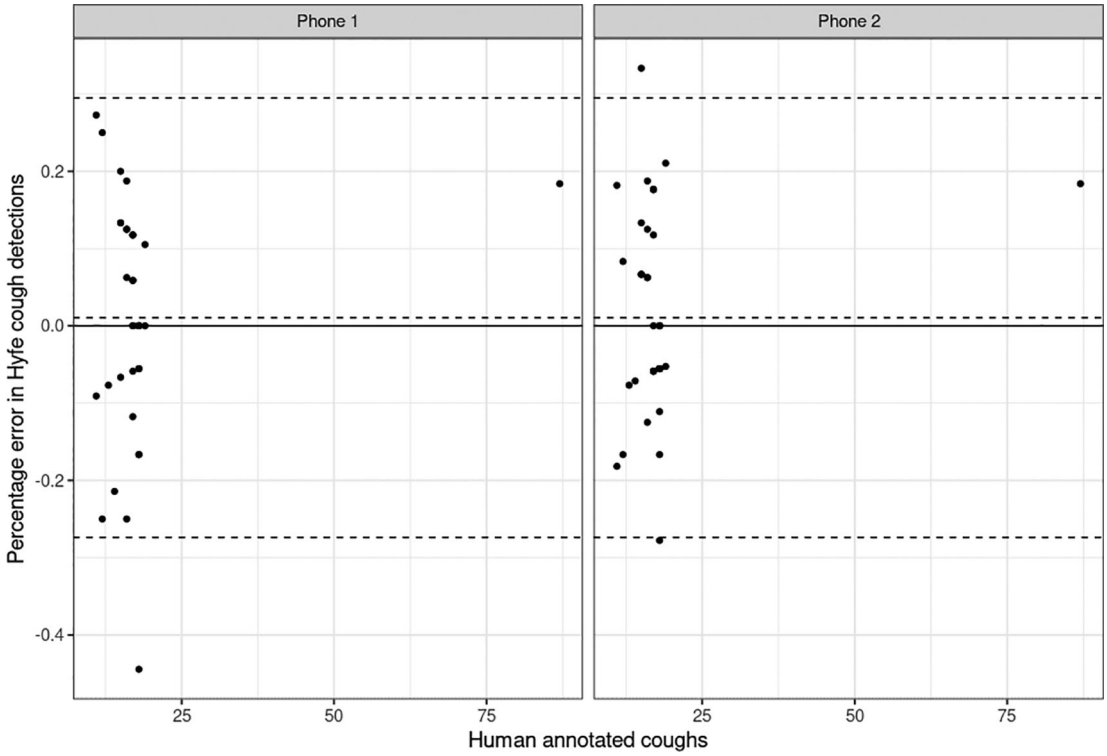

Figure 6. **Bland-Altman plot of Human annotated coughs and Percentage error in Hyfe cough detections.**

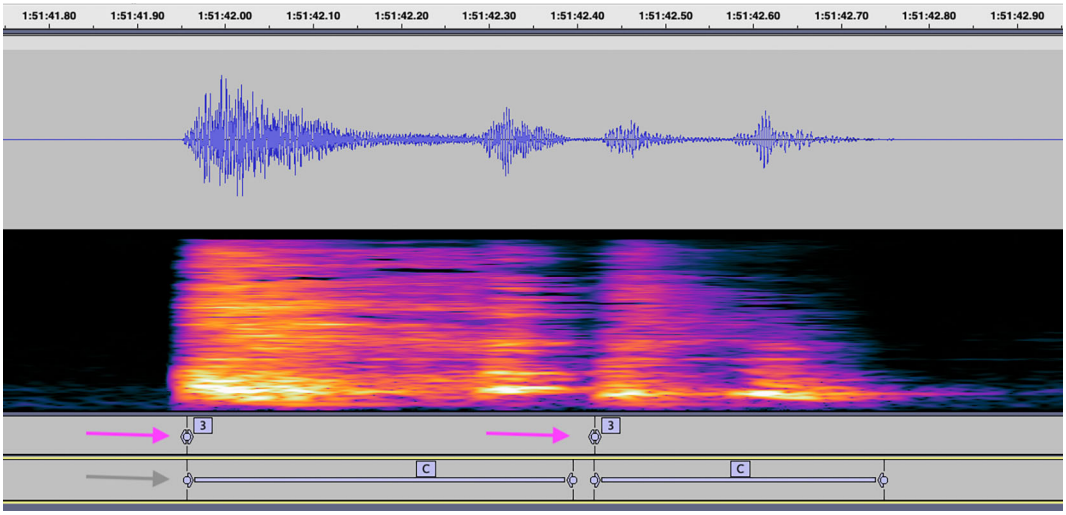

Figure 7. Example labels indicating cough placed by a human listener in a single second. Purple arrows indicate labels placed in this study, according to the 4-tier SOP. The gray arrow indicates how this audio segment would be annotated using the updated 6-tier SOP.

of such sounds may represent a challenge in the first place, leading to either unrecorded coughs or recorded and misclassified non-cough sounds. We previously reported on the analytical performance of Hyfe.[13] Here we report on its pre-clinical performance using scripted solicited coughs in a controlled environment.

Defining a gold standard for the performance evaluation of passive cough monitors represents a challenge which we addressed with standardized procedures ensuring human listener inter-observer consensus. This process and our results highlight three important issues related to evaluating cough monitors. Firstly, it is critical to have a precise method of aligning different data streams. Our failure to have this resulted in the exclusion of two sessions. Going forward, we propose the use of a distinct auditory signal, or "coda", that can be played at the beginning of each session so both the continuous audio recorder and the smartphone running the app will have a series of characteristic peak sounds that can be used for timestamping and alignment. The coda currently used for Hyfe-related studies is available on YouTube. Secondly, although solicited coughs have been used to validate cough-counting devices in the past[20] and previous literature reports that spontaneous and solicited coughs have similar acoustic characteristics,[21] we found significant differences in the sound of solicited coughs from different study participants. When asked to voluntarily cough, ten of the 49 research subjects generated sounds that were not unanimously recognized as coughs by human annotators. This observation raises questions about the utility of solicited coughs for diagnostic purposes. Finally, there are interpersonal differences in how sounds are classified by annotators. Because of this we had to exclude 88 sounds from the analysis. This has prompted additional efforts to minimize interobserver variability by developing clear operating procedures and training programs for cough annotators. We propose that protocols such as these be shared, and that consensus be sought so as to facilitate comparison of monitoring technology.

As convolutional neural networks are employed by Hyfe – they learn by example. As long as the training data is relatively unbiased and representative, a neural net can identify a "feature" (such as the acoustic signature of a cough) in a myriad of samples, even if those samples do not resemble each other. After this study, we believe that labeling cough duration rather than just its beginning has more value in further training Hyfe's AI model, also in analyzing agreement between human listeners, and agreement with Hyfe (Figure 7). Therefore, the updated 6-tier version SOP was proposed, which is currently being used for cough labeling in continuous audio recordings.

Environmental sounds may interfere with capturing coughs in real life, as seen in the sensitivity of session 38 (Figure 3), however, continuous improvements of the AI peak detection models and the cough classifiers, may address this potential issue in the near future. Even though we have not observed any significant differences in the quality of smartphones used in this trial, there might be cases when the version of smartphone operating system plays some role in smartphone's general usability and experience for the user.

Overcoming these challenges, we were able to evaluate Hyfe's accuracy using 1679 solicited sounds generated by a total of 37 subjects. Hyfe's overall sensitivity and specificity were respectively 91% and 98% and did not differ significantly between two phones. Importantly, we feel the more relevant parameter of performance to be the Pearson correlation of the

cough rates as measured by the device and the gold standard (human annotation), which was 0.968. We propose that going forward, analysis of cough monitors should use correlation in rates (gold standard vs monitor detections) as the primary metric of their performance. Though we used a minute (due to the highly condensed nature of the study), in most continuous monitoring use cases, coughs per *hour* is likely to be the most clear and useful period of observation.

Of note, the performance was lower in four subjects, presumably due to the intrinsic acoustic characteristics of solicited coughs and the level of background noise.

Our own data from more than 400 hours monitoring multiple patients with respiratory diseases in real-world environments shows a clear correlation between total coughs and cough seconds – this work is being prepared for publication. We are also analysing cough-seconds and the notion of bouts in the continued work. In the meantime, the objective of this work was to analyse the performance in detecting sounds, capturing and classifying coughs from solicited sounds in a controlled environment.

Further validation studies will need to be conducted in the specific clinical settings in which Hyfe is intended to be used. To better contextualize and design such trials, target product performance specifications will be required and are expected to differ significantly between use cases. Lessons can be learned from other types of monitors such as fitness trackers, whose results can differ from each other by up to 30%.[22] Whereas, regulated medical devices used in clinical practice will require greater precision. The presented data here is encouraging, suggesting that Hyfe's performance is adequate to proceed to validation in clinical context. Taken together, these results show that AI-enabled systems might provide a valuable tool for objectively, and unobtrusively monitoring cough.

Data availability

Underlying data

Github: hyfe-ai/navarra_performance, https://doi.org/10.5281/zenodo.7936608.[23]

This project contains the following R scripts and data:

- 01.results. R (takes pre-formatted datasets and carries out performance evaluation, plots results)
- detections.csv (Hyfe detections data)
- hyfe_performance.R (analysis of Hyfe performance)
- labels.csv (human labeled data)
- offsets_emk.csv (automatic and manual offsets made to the data)

Software

Software available from:

R version 2.04 (RStudio Team, 2020), available from https://cran.r-project.org/bin/windows/base/old/4.0.2/

Hyfe, version acl 1.24.4, available from https://www.hyfe.ai/

Audacity | Free, open source, cross-platform audio software for multi-track recording and editing, available from https://www.audacityteam.org/

References

1. Cornford CS: **Why patients consult when they cough: a comparison of consulting and non-consulting patients.** *Br. J. Gen. Pract.* [Internet] Royal College of General Practitioners; 1998 [cited 2022 Jan 19]; **48**: 1751–1754.
PubMed Abstract | Free Full Text

2. Motulsky A, Weir DL, Liang MQ, *et al.*: **Patient-initiated consultations in community pharmacies.** *Res. Soc. Adm. Pharm.* 2021 [cited 2022 Jan 19]; **17**: 428–440. Elsevier Inc.
PubMed Abstract

3. World Health Organization (WHO): **WHO operational handbook on tuberculosis. Module 2: screening - systematic screening for tuberculosis disease. Modul. 3 Diagnosis Rapid diagnotics Tuberc. diagnosis.** 2021 [cited 2022 Jan 19].
Reference Source

4. Tashkin DP, Volkmann ER, Tseng CH, *et al.*: **Improved Cough and Cough-Specific Quality of Life in Patients Treated for Scleroderma-Related Interstitial Lung Disease: Results of Scleroderma Lung Study II.** *Chest.* 2017; **151**: 813–820. Elsevier Inc.
PubMed Abstract | Publisher Full Text

5. McCallion P, De Soyza A: **Cough and bronchiectasis.** *Pulm. Pharmacol. Ther. Pulm Pharmacol Ther.* 2017 [cited 2022 Jan 19]; **47**: 77–83.
Publisher Full Text | Reference Source

6. Kvapilova L, Boza V, Dubec P, *et al.*: **Continuous Sound Collection Using Smartphones and Machine Learning to Measure Cough.** *Digit. Biomarkers.* 2019; **3**: 166–175. S. Karger AG.
PubMed Abstract | Publisher Full Text

7. Irwin RS: **Assessing cough severity and efficacy of therapy in clinical research: ACCP evidence-based clinical practice guidelines.** *Chest.* 2006 [cited 2022 Jan 19]; **129**: 232S–237S.
PubMed Abstract | Publisher Full Text

8. Birring SS, Fleming T, Matos S, *et al.*: **The Leicester Cough Monitor: Preliminary validation of an automated cough detection system in chronic cough.** *Eur. Respir. J.* 2008 [cited 2021 Dec 22]; **31**: 1013–1018.
Publisher Full Text | Reference Source

9. Crooks MG, Hayman Y, Innes A, *et al.*: **Objective Measurement of Cough Frequency During COPD Exacerbation Convalescence.** *Lung.* 2016 [cited 2021 Dec 22]; **194**: 117–120.
PubMed Abstract | Publisher Full Text

10. Soliń ski M, Łepek M, Kołtowski Ł: **Automatic cough detection based on airflow signals for portable spirometry system.** *Informatics Med. Unlocked.* 2020; **18**: 100313. Elsevier.
Publisher Full Text

11. Porter P, Abeyratne U, Swarnkar V, *et al.*: **A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children.** *Respir. Res.* 2019 [cited 2022 Jan 19]; **20**: 81.
PubMed Abstract | Publisher Full Text

12. Gabaldón-Figueira JC, Keen E, Rudd M, *et al.*: **Longitudinal passive cough monitoring and its implications for detecting changes in clinical status.** *ERJ Open Res.* 2022 May 16; **8**(2): 00001–02022.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Gabaldon-Figueira JC, Brew J, Doré DH, *et al.*: **Digital acoustic surveillance for early detection of respiratory disease outbreaks in Spain: A protocol for an observational study.** *BMJ Open.* 2021 [cited 2022 Jan 19]; **11**: 51278.
PubMed Abstract | Publisher Full Text

14. Gabaldón-Figueira JC, Keen E, Giménez G, *et al.*: **Acoustic surveillance of cough for detecting respiratory disease using artificial intelligence.** *ERJ Open Res.* 2022; **8**: 00053–02022. in press.
PubMed Abstract | Publisher Full Text

15. Audacity® | Free, open source, cross-platform audio software for multi-track recording and editing: [cited 2022 Jan 24].
Reference Source

16. Smith JA, Earis JE, Woodcock AA: **Establishing a gold standard for manual cough counting: video versus digital audio recordings.** *Cough.* 2006 Aug 3; **2**: 6.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Lake C, Briffa P, Munoz P, *et al.*: **Documentation of cough provoked during a mannitol challenge using acoustic respiratory monitoring compared to video surveillance monitoring [Conference Abstract].** *Respirology.* 2012; **17**: 1.

18. Hajian-Tilaki K: **Sample size estimation in diagnostic test studies of biomedical informatics.** *J. Biomed. Inform.* 2014; **48**: 193–204. Academic Press.
PubMed Abstract | Publisher Full Text

19. Arfin WN: **wnarifin.github.io > Sample size calculator.** *Sample size Calc.* 2021 [cited 2022 Jan 19].
Reference Source

20. Vizel E, Yigla M, Goryachev Y, *et al.*: **Validation of an ambulatory cough detection and counting application using voluntary cough under different conditions.** *Cough Bio. Med. Central.* 2010 [cited 2022 Jan 19]; **6**: 1–8.
Publisher Full Text

21. Korpáš J, Sadločová J, Vrabec M: **Analysis of the cough sound: An overview.** *Pulm. Pharmacol. Pulm Pharmacol.* 1996 [cited 2022 Jan 19]; **9**: 261–268.
Publisher Full Text | Reference Source

22. Jagim AR, Koch-Gallup N, Camic CL, *et al.*: **The accuracy of fitness watches for the measurement of heart rate and energy expenditure during moderate intensity exercise.** *J. Sports Med. Phys. Fitness.* 2021 [cited 2022 Jan 19]; **61**: 205–211.
Publisher Full Text | Reference Source

23. Galvosas M, Gabaldón-Figueira JC, Keen EM, *et al.*: **Performance evaluation of the smartphone-based AI cough monitoring app - Hyfe Cough Tracker against solicited respiratory sounds.** *F1000 Research.* 2023.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✗

Version 1

Reviewer Report 08 September 2022

https://doi.org/10.5256/f1000research.134609.r148456

✗

Terence E. Taylor
[1] Vitalograph Ireland Ltd., Ennis, Ireland
[2] Vitalograph Ireland Ltd., Ennis, Ireland

Editorial note [4th May 2023]:
A potential conflict of interest has come to light and so we have added this to the report to ensure full transparency.

The aim of the Hyfe app is to provide a portable and high performing system for measuring cough rate. The authors describe an evaluation of the Hyfe app using 10-minute audio recordings consisting of cough and speech obtained from participants in a controlled environment. Audio data were collected using two identical smartphones and an additional MP3 recorder placed 50cm to participants.
While I find this area of research quite interesting, I have several major concerns with this study from the data collection to how the evaluation was measured. The type and amount of data does not reflect how the app would be used in real-life settings. In my opinion, this study does not provide sufficient evidence that the Hyfe system can perform adequately in real-life clinical research, particularly over longer periods of time (24 hours or above). Furthermore, I believe this study lacks novelty compared to other audio-based cough detection studies and represents more of a pilot study than an evaluation of the system. Please see further comments below.

Abstract:

- In Objectives, please change the relevant sentence to “We evaluate performance of cough detection using continuous audio recordings obtained within a controlled environment and cough counting by trained individuals as the gold standard.”

Introduction:

- MAJOR: In ref 13, it is noted that participants in this other study complained of increased battery consumption. And so, the evaluation there was constrained to 6-hour windows. Has this app been ever evaluated over at least 24-hour periods? According to the authors, one of the advantages of Hyfe is the ability to passively record for extended periods of time (i.e.

greater than 24 hours). Has there been analysis done on the battery consumption effects over longer periods? In my opinion, this study should have monitored cough over at least 24 hours to show a more realistic evaluation of the system.

- MINOR: Privacy concerns are mentioned in the introduction. How does Hyfe preserve privacy exactly? If this information has already been published elsewhere, please cite the relevant literature and give a brief explanation on how privacy preservation is achieved.

Methods:

- MINOR: What versions of iOS/Android has the Hyfe app been tested on?

- MINOR: Why were baseline respiratory symptoms not considered for participants? Cough sounds can vary in terms of acoustic properties across different respiratory disease types. The data reported here have no mention of being evaluated on any respiratory disease cough sounds.

- MINOR: Regarding the CNN model, was it trained on sufficient data across different smartphones, microphone locations, disease types? Please comment on this.

- MAJOR: Why place the smartphone 50cm away from the participant? What is the approximate distance between a user's mouth and a smartphone mic when they are interacting with their smartphone in front of them with smartphone-in-hand? Was this thought of to replicate real-life use? I would have thought that in many real-life cases also, a lot of the audio that Hyfe would capture would be when the smartphone is located inside the user's pocket or bag. Why not place one smartphone in the participant's pocket and have one in their hand to try replicate real-life use? Why constrain the recording protocol so heavily? Also, why use two of the same identical smartphones? It would have been useful to compare different smartphones with different builds, battery consumption properties, microphone designs etc. Was the gain/dynamic range of the microphones taken into account? All these points significantly limit the interpretability of this evaluation in relation to how Hyfe could perform in clinical settings in my opinion. The data collection is not representative of real-life use.

- MINOR: What is the sampling rate and bit depth of the audio recordings? Please report this.

- MINOR: It would be useful to have an English language version of the pre-generated computer script also available to the reader.

- MINOR: How were participants instructed to cough exactly? Was it for a certain amount of time? For a certain number of explosive cough sounds? Please explain briefly.

- MAJOR: Recording sessions occurred in a quiet room as reported. What does "quiet" mean here? Were there sound intensity levels recorded? Was the room acoustically insulated? What were the signal-to-noise ratio (SNR) values of the different sessions? Please provide this information to help the reader understand the environment the recordings were obtained in.

- MAJOR: What were the kappa agreement scores to show inter-rater agreement for the

manual labeling? Why do you need a 4-tier labeling system when users were prompted to perform specific sounds? Do you not already know when a user coughed as they were prompted to do so at specific points within the 10min session? I understand the prompts were randomized, but was this not recorded internally for post-analysis? Therefore, why is there a need for labels 1 and 2 if you have a log of what the participants were doing during each part of the recordings?

- MINOR: How many sounds were labeled 0, 1, 2 and 3? Please report this information. How many explosive cough sounds on average were there in each cough audio segment? Were participants prompted to perform a certain number of explosive cough sounds?

- MINOR: Why does the Hyfe system classify every 0.5 seconds, but labeling was done every 1 second?

- MAJOR: Multiple coughs may occur within a 1 second window. Why label a full 1 second window as a single true positive cough even if there could be multiple single manually labeled coughs within the 1 second? In my opinion, cough rate estimation would be more accurate if all coughs were considered rather than labeling a 1 second window of coughing as a single cough event (when estimating app performance). Is there reasoning for this in the literature? If so, please add relevant citations. If I understand correctly, it means technically you could miss all but one cough in a given 1sec window and still have very high sensitivity/specificity based on your labeling system? If this is the case, this is a major limitation. How coughs are grouped together in peals or bouts can have a significant impact on cough rate estimation. I think this discrepancy would become more evident had the system been evaluated on a cough-by-cough basis and over longer periods on patients with respiratory disease as they may tend to cough in groups in quick succession (within 1 second in many cases).

- MINOR: Was there any overlap in the labeling of 1 second windows i.e. 50% overlap between consecutive windows for labeling?

- MAJOR: Sample Size: “385 sounds” – are you referring to cough sounds here, or a mix of both cough and non-cough? How did you determine the balance of data between cough and non-cough (speech, throat clear, noise etc.)? It seems a balance of approximately 1:2 (coughs to non-coughs) was employed. In real-life settings, over the course of a 24-hour period for example, the number of non-cough segments will far outweigh the number of cough segments if Hyfe was to continuously record audio. This does not seem to be considered here. If the data were to be more realistically balanced in this way, it may have a significant effect on the positive predictive value (PPV) of the algorithm (PPV needs to be reported). One of the biggest challenges of longer term (24 hours or more) audio-based cough monitoring is the PPV performance measure.

- MAJOR: One of the advantages of Hyfe, reported by the authors, is that it can monitor cough rate over longer periods (>24 hours). But this study looked at 10-minute recording windows. Even taking into account that no respiratory symptoms were not noted from participants (which I think is a major limitation), I think the balance of data here are not truly representative of real-life settings especially if the authors suggest Hyfe can monitor cough rate over 24-hour periods. This is a major limitation to this evaluation study. This

system needs to be evaluated over at least 24-hour periods before it can be suggested that it may perform strongly for longer term analysis in clinical settings.

- MINOR: Synchronization of labels - Please provide a clearer explanation for this. Why synchronize to within 2 seconds? This seems like a long duration considering multiple explosive cough sounds may occur within 2 seconds.

- MINOR: Data were discarded if they had less than 10 cough seconds of agreement. How often did that occur? How much data were discarded overall? – from Figure 1 it seems 20-25% of data were discarded due to inadequate quality. How would this reflect real-life evaluation then? The app would be unsure of about 20% of coughs at best? That seems to be a potential large margin of error. Please comment on this.

- MAJOR: Was the 4-tier or 6-tier labeling SOP used? This is quite confusing for the reader. If 6-tier is superior in the opinion of the authors, why not revert and apply it to all data and re-run the analysis? I understand it may mean re-training the CNN but if it will be beneficial as mentioned by the others, then it should be reported here.

- MINOR: "Because the utility of a cough monitor is not in noting individual coughs but rather in tracking cough rates…". Surely identifying singular coughs should be the goal to obtain the most accurate estimation of cough rate? If not, you could hypothetically be detecting false positives and false negatives, but still hit the correct cough rate. This could affect the analysis of how cough rate changes over time throughout a 24-hour period, for example, if it were of interest. Please provide a stronger argument for this.

Results:

- MAJOR: Please report SNR for audio sessions in comparison to cough. It may help the reader understand the poor performance of sessions 2, 17 and 38. Real-life recordings will be full of different types and levels of noise. If it is only possible to evaluate on audio within controlled environments, then the authors should augment the data by adding various different types and levels of noise to the data to replicate real-life use.

- MINOR: What were the uncommon acoustic characteristics associated with Sessions 2 and 17? An additional figure with time domain plot as well as a spectrogram plot of coughs from these participants would be useful here for the reader to understand why the system did not perform as well for these participants.

- MINOR: Figure 3 image quality should be improved.

- MINOR: "Sensitivity for the Session 20 was not evaluated because this was a patient with refractory chronic cough that generated hundreds of out-of-script, making timestamping impossible." – Please elaborate further. Were there too many coughs? One would assume the data from this participant would have been very interesting to analyze and run through the Hyfe system?

- MINOR: Limitations: I agree with the limitations mentioned. So why evaluate the Hyfe system in a way it will rarely if not never be used in? Please comment on this.

- MAJOR: Please give details of the effect the Hyfe app has on battery life. One would think a continuous audio recording will significantly drain a smartphone battery. It seems to be mentioned in a previous study (ref 13). Were there experiments done on this? If not, this analysis needs to be performed and reported in this evaluation study.

- MAJOR: A Bland Altman plot analyzing cough rate between Hyfe and manual labels is required here. Correlation analysis is not enough to convince the reader that the Hyfe system can accurately estimate cough rate (unless you were to use the linear regression equation to map Hyfe estimation to manual estimation which I don't think is the goal here). While strong correlation is interesting to observe, simply showing the equal line on Figure 4 does not highlight if there is any bias in cough rate estimation. Please add the linear regression line and equation in Figure 4. Please also add an additional panel for a Bland Altman plot comparing the mean cough rate between manual and Hyfe to the difference in cough rate between manual and Hyfe.

Discussion:

- MAJOR: Where does the score threshold of 0.85 come from? Can you report the distribution of CNN output probability scores for all label categories as supplementary material? A histogram, for example, of probability scores for each label category would be interesting to see and could enhance the reader's understanding of the challenges in labeling category 1 and 2 sounds.

- MINOR: The discussion on inter-participant variability and the debate on using solicitated voluntary coughs as a diagnostic tool is quite interesting. Does this suggest these are most likely voluntary throat clears rather than involuntary coughs? As a result, why then prompt the user to record a voluntary cough when first using the app?

- MINOR: What difference will the 6-tier SOP make for training the CNN model? I assume fixed 0.5sec segments are employed for training the model so how will the duration markers be more beneficial than just the onset markers used in the 4-tier SOP labeling system?

- MAJOR: The Pearson correlation is interesting, but I disagree that it should be the primary metric of performance. I think the authors should perform a linear regression analysis also to analyze the slope of the regression line. Only using Pearson correlation does not disclose information regarding how many coughs are missed in detection. For example, you could theoretically have a Pearson coefficient of 1 but you are missing 1 cough in every 2 or 3 coughs consistently in your detection model (this would be reflected in the slope of the regression analysis). The Pearson correlation metric is one of a group of metrics that should be considered in my opinion. I would like to see linear regression analysis, absolute percentage error according to the total number of coughs per participant and a Bland Altman analysis.

- MAJOR: "...Hyfe's performance is adequate to be used in most clinical and research contexts." This study does not suggest this in my opinion if the data were collected in controlled environments, was not evaluated on real respiratory patients and the smartphones were not evaluated in multiple different positions/locations.

Other Comments:

- MAJOR: While I understand the data were obtained in controlled environments, in my opinion, more data are required to obtain a more realistic measure of performance in order to suggest Hyfe could be used in clinical settings. This includes more data using different smartphone locations (in user's pocket, bag etc.), more realistic data with different types and levels of noise (including muffled coughs due to hand/mask covering) and I think it should be evaluated on a cough-by-cough basis rather 1 second windows as a form of cough rate.

- MINOR: How does the app know if it is the user who coughed and not a person beside them in a crowded area? Has this been taken into account within the CNN training data? Please comment on this.

- MINOR: Does the Hyfe app still record audio when the user is on a phone call or sending voice messages?

- MINOR: Does the app work without internet connectivity?

- MINOR: SOP4 Supplementary Document: Page 4, Section 3, bullet 3 – should that read as "…at the end of the expulsive phase.."?

- MINOR: SOP4 Supplementary Document: I think reporting a cough as having one sound or two sounds is a little confusing. Consider moving Fig. 5 and the paragraph above it to before Fig. 3. Start with showing the sound, explaining the three phases of the cough sound event and then describe in more detail the physiology behind it. Fig. 3 panel B needs to highlight that the voiced phase is missing. It is a little confusing to just say A has a "double" sound and B has a "single sound".

- MAJOR: SOP4 Supplementary Document: Page 8, Table 1 – according to the 4-tier process, the label 2 may be due to the phone located in the user's pocket which, I assume, will be a vastly common occurrence if this app were to be used in cough clinical applications. In this study label 2 coughs weren't considered in the app evaluation. Why not? I think it is essential particularly for an app-based cough monitoring system to be evaluated on recordings when the phone is in the user's pocket or bag for example.

- MINOR: SOP4 Supplementary Document: Page 8, Labeling Tips – I think the label 2 may highlight a potential limitation of the Hyfe system. It is indicating to the reader that the smartphone needs to be in close proximity to the user. Coughs may sound "distant" even if the user is coughing. Please elaborate what "distant" means here. Is this related to the audio amplitude of the cough sound? Certain types of coughs may have lower amplitudes, particularly if a very ill patient has exhausted their respiratory strength.

- MAJOR: The smartphone can be located on a table with many other people in the user's vicinity. How do you know it is the user's cough? This could be a major limitation. Particularly if the intention is to detect subtle changes in cough rate. There are many external factors when using an app-based audio cough detection system that can affect cough rate estimation. I don't think the authors make it clear how they will evaluate this or can overcome such challenges using this system.

- MINOR: If possible, it would be very beneficial to have examples of the sounds that were recorded in the controlled environment for readers to listen to.

Is the work clearly and accurately presented and does it cite the current literature?
No

Is the study design appropriate and is the work technically sound?
No

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: I am an employee of Vitalograph (Ireland) Ltd but this has not affected my ability to review impartially.

Reviewer Expertise: audio signal processing, cough sound analysis, artificial intelligence, telehealth, respiratory diseases

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 27 Apr 2023

Mindaugas Galvosas

2.1
Comment: While I find this area of research quite interesting, I have several major concerns with this study from the data collection to how the evaluation was measured. The type and amount of data does not reflect how the app would be used in real-life settings. In my opinion, this study does not provide sufficient evidence that the Hyfe system can perform adequately in real-life clinical research, particularly over longer periods of time (24 hours or above). Furthermore, I believe this study lacks novelty compared to other audio-based cough detection studies and represents more of a pilot study than an evaluation of the system. Please see further comments below.

Answer: Firstly we want to raise a serious issue with this review. Terence Taylor is a Senior

Clinical Data Scientist working for Vitalograph Ireland Ltd, a for profit company that manufactures the Vitalojak, a device used to quantify cough in 24 hour windows in the context of clinical trials. This product could be affected by the emergence of new technologies capable of monitoring cough for longer periods of time. There is a clear economic conflict of interest and the reviewer should have either recused himself or clearly disclosed it. None of these potential remedial actions were taken.

We find it equally appalling that the editorial team of F1000 research has failed to act on this after the issue was raised.

Here we provide point by point answers to the reviewer´s comments assuming good will, even when the repeated demands for experiments and data outside the scope of this manuscript clearly suggest otherwise.

This was the laboratory evaluation of our algorithm, and this manuscript is not meant to demonstrate how it performs in real world environments. Such evaluation, which is a necessary next step, is being done and will be provided as a separate manuscript showing real-world validation of the technology. The scope of this manuscript did not include real-life scenarios and long-term monitoring.

The novelty of this study is purely in evaluating the performance of novel fully automated sound capture, cough detection and cough quantification system.

Actions: None taken.

2.2
Comment:
Abstract: In Objectives, please change the relevant sentence to “We evaluate performance of cough detection using continuous audio recordings obtained within a controlled environment and cough counting by trained individuals as the gold standard.”

Answer: Abstract edited as per reviewer recommendation.

Actions: Abstract edited as per reviewer recommendation.

2.3
Comment: Introduction: MAJOR: In ref 13, it is noted that participants in this other study complained of increased battery consumption. And so, the evaluation there was constrained to 6-hour windows. Has this app been ever evaluated over at least 24-hour periods? According to the authors, one of the advantages of Hyfe is the ability to passively record for extended periods of time (i.e. greater than 24 hours). Has there been analysis done on the battery consumption effects over longer periods? In my opinion, this study should have monitored cough over at least 24 hours to show a more realistic evaluation of the system

Answer: This question is not related to the experiment described in the manuscript, which was purely evaluating the performance of the algorithm. “...one of the advantages of Hyfe is

the ability to passively record for extended periods of time (i.e. greater than 24 hours)." – we have conducted separate works that followed-up for long periods of time (months) and that is available: (reference 12, doi: 10.1183/23120541.00001-2022; also reference 14, https://doi.org/10.1183/23120541.00053-2022)

Actions: References of long-term monitoring are reflected (12, 14).

2.4
Comment: Introduction: MINOR: Privacy concerns are mentioned in the introduction. How does Hyfe preserve privacy exactly? If this information has already been published elsewhere, please cite the relevant literature and give a brief explanation on how privacy preservation is achieved.

Answer: Privacy concerns are described in previous publications (references of this manuscript 12-14) that include real patients and monitoring in real world environments. The scope of this manuscript was not to overview the whole monitoring process, but to evaluate the performance of the technology in controlled environments with solicited coughs.

Actions: Manuscript references 12-14 address privacy concerns and describe privacy preserving in detail.

2.5
Comment: Methods: MINOR: What versions of iOS/Android has the Hyfe app been tested on?

Answer: This study was done with Android phones only and all used Android version 11.

Actions: Android version 11 is mentioned in the manuscript.

2.6
Comment: Methods: MINOR: Why were baseline respiratory symptoms not considered for participants? Cough sounds can vary in terms of acoustic properties across different respiratory disease types. The data reported here have no mention of being evaluated on any respiratory disease cough sounds.

Answer: Here again, the objective of this work was to evaluate the performance on solicited coughs. There is ongoing work validating the algorithm in clinical patients in real-life environments.

Actions: None taken.

2.7
Comment: Methods: MINOR: Regarding the CNN model, was it trained on sufficient data across different smartphones, microphone locations, disease types? Please comment on this.

Answer: It was done with a broad range of microphones and collected from various

locations and it has been added to the manuscript.

Actions: Added to "Methods": "Hyfe's CNN model, at the time of this analysis, was trained on more than 200M real-world cough and cough-like samples, collected from multiple countries and multiple devices."

2.8
Comment: Methods: MAJOR: Why place the smartphone 50cm away from the participant? What is the approximate distance between a user's mouth and a smartphone mic when they are interacting with their smartphone in front of them with smartphone-in-hand? Was this thought of to replicate real-life use? I would have thought that in many real-life cases also, a lot of the audio that Hyfe would capture would be when the smartphone is located inside the user's pocket or bag. Why not place one smartphone in the participant's pocket and have one in their hand to try replicate real-life use? Why constrain the recording protocol so heavily? Also, why use two of the same identical smartphones? It would have been useful to compare different smartphones with different builds, battery consumption properties, microphone designs etc. Was the gain/dynamic range of the microphones taken into account? All these points significantly limit the interpretability of this evaluation in relation to how Hyfe could perform in clinical settings in my opinion. The data collection is not representative of real-life use.

Answer: Here again, you are requesting information which goes beyond the scope of the work described in the manuscript. The main objective of this work was to evaluate the performance (sound capture, cough detection and quantification) of the AI model on solicited coughs in a controlled environment.

Actions: None taken.

2.9
Comment: Methods: MINOR: What is the sampling rate and bit depth of the audio recordings? Please report this.

Answer: The sampling rate was 44.1 Hz and the files are 16-bit.

Actions: Added to study design: "The sampling rate was 44.1 Hz and the files are 16-bit."

2.10
Comment: Methods: MINOR: It would be useful to have an English language version of the pre-generated computer script also available to the reader.

Answer: We fail to see the additional value of a translation, given that the Spanish words were chosen based on their pronunciation and cough-like sounding.

Actions: None taken.

2.11
Comment: Methods: MINOR: How were participants instructed to cough exactly? Was it for

a certain amount of time? For a certain number of explosive cough sounds? Please explain briefly.

Answer: This is a valid comment. Participants were instructed to cough once every time they were prompted by the script to do so.

Actions: Added "Participants were instructed to cough once every time they were prompted by the script to do so." to "Study design".

2.12
Comment: Methods: MAJOR: Recording sessions occurred in a quiet room as reported. What does "quiet" mean here? Were there sound intensity levels recorded? Was the room acoustically insulated? What were the signal-to-noise ratio (SNR) values of the different sessions? Please provide this information to help the reader understand the environment the recordings were obtained in.

Answer: This is a valid comment. Sound intensity levels were monitored using https://meters.uni-trend.com/product/ut353-ut353bt/ and never exceeded 110dB.
The room was not acoustically insulated.

Actions: Added to the "Study design": "Sound intensity levels in the room were also monitored using UNI-T mini sound level meter. The room was not acoustically insulated". Added to the "Results": "Sound levels in the room during the study were never above 110dB."

2.13
Comment: Methods: MAJOR: What were the kappa agreement scores to show inter-rater agreement for the manual labeling? Why do you need a 4-tier labeling system when users were prompted to perform specific sounds? Do you not already know when a user coughed as they were prompted to do so at specific points within the 10min session? I understand the prompts were randomized, but was this not recorded internally for post-analysis? Therefore, why is there a need for labels 1 and 2 if you have a log of what the participants were doing during each part of the recordings?

Answer: The sounds were indeed randomized and triple-human labeling was used as the gold standard. The nature of an elicited sound, even though the instruction is to produce a cough, varies depending on the person (some may get confused and do throat clears when asked to cough, etc).
We agree that inter-rater agreement is an interesting point to look at. There is a separate full paper in preparation looking specifically at this with a whole different dataset.

Actions: Added a linear analysis plot and Bland-Altman plot in the "Results" section.

2.14
Comment: Methods: MINOR: How many sounds were labeled 0, 1, 2 and 3? Please report this information. How many explosive cough sounds on average were there in each cough audio segment? Were participants prompted to perform a certain number of explosive

cough sounds?

Answer: the dataset has been made public with this manuscript. The number of sounds instructed was the same for all participants, as prompted by the script.

Actions: Added to "Study design": "In total for each participant, the script included instructions to cough 20 times, sneeze 10 times, clear throat 5 times and produce 15 sounds (explosive words, for example, "paella" and numbers as "93")."

2.15
Comment: Methods: MINOR: Why does the Hyfe system classify every 0.5 seconds, but labeling was done every 1 second?

Answer: This is a misinterpretation. Hyfe uses 0.5s snippets to classify explosive sounds. Labelers were instructed to label every beginning of a cough sound, and the unit of analysis chosen was cough seconds.

Actions: None taken.

2.16
Comment: Methods: MAJOR: Multiple coughs may occur within a 1 second window. Why label a full 1 second window as a single true positive cough even if there could be multiple single manually labeled coughs within the 1 second? In my opinion, cough rate estimation would be more accurate if all coughs were considered rather than labeling a 1 second window of coughing as a single cough event (when estimating app performance). Is there reasoning for this in the literature? If so, please add relevant citations. If I understand correctly, it means technically you could miss all but one cough in a given 1sec window and still have very high sensitivity/specificity based on your labeling system? If this is the case, this is a major limitation. How coughs are grouped together in peals or bouts can have a significant impact on cough rate estimation. I think this discrepancy would become more evident had the system been evaluated on a cough-by-cough basis and over longer periods on patients with respiratory disease as they may tend to cough in groups in quick succession (within 1 second in many cases).

Answer:
Our own data from more than 400hrs monitoring multiple patients with respiratory diseases in real-world environments shows a clear correlation between total coughs and cough seconds – this work is being prepared for publication. We are also analysing cough-seconds and the notion of bouts in the continued work. In the meantime, the objective of this work was to analyse the performance in detecting sounds, capturing and classifying coughs from solicited sounds in a controlled environment.

Actions: Added to the "Discussion": "Our own data from more than 400hrs monitoring multiple patients with respiratory diseases in real-world environments shows a clear correlation between total coughs and cough seconds – this work is being prepared for publication. We are also analysing cough-seconds and the notion of bouts in the continued work. In the meantime, the objective of this work was to analyse the performance in

detecting sounds, capturing and classifying coughs from solicited sounds in a controlled environment."

2.17
Comment: Methods: MINOR: Was there any overlap in the labeling of 1 second windows i.e. 50% overlap between consecutive windows for labeling?

Answer: No, there was not.

Actions: None taken.

2.18
Comment: Methods: MAJOR: Sample Size: "385 sounds" – are you referring to cough sounds here, or a mix of both cough and non-cough? How did you determine the balance of data between cough and non-cough (speech, throat clear, noise etc.)? It seems a balance of approximately 1:2 (coughs to non-coughs) was employed. In real-life settings, over the course of a 24-hour period for example, the number of non-cough segments will far outweigh the number of cough segments if Hyfe was to continuously record audio. This does not seem to be considered here. If the data were to be more realistically balanced in this way, it may have a significant effect on the positive predictive value (PPV) of the algorithm (PPV needs to be reported). One of the biggest challenges of longer term (24 hours or more) audio-based cough monitoring is the PPV performance measure.

Answer: As the text states, this refers to the total number of sounds. Once more, this study did not evaluate the balance of cough and non-cough segments over a long period of time, as the objective was to only evaluate the performance of automated listening to elicited sounds, capturing and quantifying elicited coughs in controlled environments, which we achieved and described in detail.

Actions: None taken.

2.19
Comment: Methods: MAJOR: One of the advantages of Hyfe, reported by the authors, is that it can monitor cough rate over longer periods (>24 hours). But this study looked at 10-minute recording windows. Even taking into account that no respiratory symptoms were not noted from participants (which I think is a major limitation), I think the balance of data here are not truly representative of real-life settings especially if the authors suggest Hyfe can monitor cough rate over 24-hour periods. This is a major limitation to this evaluation study. This system needs to be evaluated over at least 24-hour periods before it can be suggested that it may perform strongly for longer term analysis in clinical settings

Answer: Here again, you are requesting data that exceeds the scope of this work, which was to evaluate the performance of the Hyfe system in controlled environments with solicited coughs. There is published data about the capacity of Hyfe to monitor patients for periods of months (references 12-14).

Actions: None taken.

2.20
Comment: Methods: MINOR: Synchronization of labels - Please provide a clearer explanation for this. Why synchronize to within 2 seconds? This seems like a long duration considering multiple explosive cough sounds may occur within 2 seconds.

Answer: We clarify that synchronization was done to the exact time of the sound. The text has been modified to reflect the sequential nature of changes.

Actions: Edits in capitalized letters: "Labels created by listeners (in Audacity) and Hyfe detected coughs were FIRST manually synchronized to within two seconds of each other. Synchronization was THEN carried out for each phone and each session separately by identifying the time offset that would align Hyfe detections with the labels and adjusting the Hyfe detection timestamps accordingly."

2.21
Comment: Methods: MINOR: Data were discarded if they had less than 10 cough seconds of agreement. How often did that occur? How much data were discarded overall? – from Figure 1 it seems 20-25% of data were discarded due to inadequate quality. How would this reflect real-life evaluation then? The app would be unsure of about 20% of coughs at best? That seems to be a potential large margin of error. Please comment on this.

Answer: See the third sentence of the "Results" section for clarification: "Ten sessions did not have at least 10 sounds unanimously labeled as coughs and were also excluded, leaving 37 sessions, with 672 unanimously-labeled cough-seconds, and 1,007 non-cough seconds for the final performance evaluation."
Additionally, the goal of this work was not to evaluate the app in real-life and a separate work - app's evaluation publication is in progress.

Actions: None taken.

2.22
Comment: Methods: MAJOR: Was the 4-tier or 6-tier labeling SOP used? This is quite confusing for the reader. If 6-tier is superior in the opinion of the authors, why not revert and apply it to all data and re-run the analysis? I understand it may mean re-training the CNN but if it will be beneficial as mentioned by the others, then it should be reported here.

Answer: 4-tier SOP was used in this study. The findings presented here informed further changes to the SOP, resulting in the 6-tier SOP. This work was described as performance evaluation of the app and its result was also the updated SOP. We will be presenting 6-tier SOP results and further updates to the SOP in other soon upcoming publications, as it was applied in multiple studies done afterwards.

Actions: None taken.

2.23
Comment: Methods: MINOR: "Because the utility of a cough monitor is not in noting

individual coughs but rather in tracking cough rates...". Surely identifying singular coughs should be the goal to obtain the most accurate estimation of cough rate? If not, you could hypothetically be detecting false positives and false negatives, but still hit the correct cough rate. This could affect the analysis of how cough rate changes over time throughout a 24-hour period, for example, if it were of interest. Please provide a stronger argument for this.

Answer: The Hyfe Cough Monitoring System recognizes and timestamps users' coughs as they occur. To manage the challenges of distinguishing coughs that happen in rapid succession (bouts), these timestamps are converted into cough-seconds as the basic unit of analysis; a cough-second is a second during which at least one cough occurs. The endpoint of our validation trial consists of hourly tabulations of these cough-seconds for each subject.

Actions: None taken.

2.24
Comment: Results: MAJOR: Please report SNR for audio sessions in comparison to cough. It may help the reader understand the poor performance of sessions 2, 17 and 38. Real-life recordings will be full of different types and levels of noise. If it is only possible to evaluate on audio within controlled environments, then the authors should augment the data by adding various different types and levels of noise to the data to replicate real-life use.

Answer: The goal of this study was not to evaluate Hyfe in real life environments. Ongoing studies validate the technology in real-life environments and publications will prove that in the near future.

Actions: None taken.

2.25
Comment: Results: MINOR: What were the uncommon acoustic characteristics associated with Sessions 2 and 17? An additional figure with time domain plot as well as a spectrogram plot of coughs from these participants would be useful here for the reader to understand why the system did not perform as well for these participants.

Answer: It is mentioned in the manuscript that sessions 2 and 17 had biphasic decibel peaks and different spectrogram features. This was a result of, for example, multiple solicited coughs being produced and the nature of the solicited sound, even though the script instructed a single cough.

Actions: None taken.

2.26
Comment: Results: MINOR: Figure 3 image quality should be improved.

Answer: "Figure 3" quality is improved.

Actions: "Figure 3" is replaced with a better quality image.

2.27
Comment: Results: MINOR: "Sensitivity for the Session 20 was not evaluated because this was a patient with refractory chronic cough that generated hundreds of out-of-script, making timestamping impossible." – Please elaborate further. Were there too many coughs? One would assume the data from this participant would have been very interesting to analyze and run through the Hyfe system?

Answer: The cough pattern of this participant has been analyzed and is described elsewhere (Reference 13).
She made an attempt to collaborate in this particular experiment as well but her chronic cough resulted in hundreds of observations outside the script.

Actions: None taken.

2.28
Comment: Results: MINOR: Limitations: I agree with the limitations mentioned. So why evaluate the Hyfe system in a way it will rarely if not never be used in? Please comment on this.

Answer: Because this was not an evaluation of the system in real world environments, but of the performance of the algorithm in a controlled environment.

Actions: None taken.

2.29
Comment: Results: MAJOR: Please give details of the effect the Hyfe app has on battery life. One would think a continuous audio recording will significantly drain a smartphone battery. It seems to be mentioned in a previous study (ref 13). Were there experiments done on this? If not, this analysis needs to be performed and reported in this evaluation study.

Answer: These details are outside the scope of this manuscript. Hyfe was successfully used by hundreds of participants in a cohort study described elsewhere with no major complaints about the battery life REF 14 in the manuscript (DOI: 10.1183/23120541.00053-2022).

Actions: None taken.

2.30
Comment: Results: MAJOR: A Bland Altman plot analyzing cough rate between Hyfe and manual labels is required here. Correlation analysis is not enough to convince the reader that the Hyfe system can accurately estimate cough rate (unless you were to use the linear regression equation to map Hyfe estimation to manual estimation which I don't think is the goal here). While strong correlation is interesting to observe, simply showing the equal line on Figure 4 does not highlight if there is any bias in cough rate estimation. Please add the linear regression line and equation in Figure 4. Please also add an additional panel for a Bland Altman plot comparing the mean cough rate between manual and Hyfe to the difference in cough rate between manual and Hyfe.

Answer: Bland-Altman plot analyzing cough rate between Hyfe and human annotators is presented as Figure 6, also Figure 5 shows the linear analysis.

Actions: Added “Figure 5” and “Figure 6” to the “Results” section for linear analysis and Bland-Altman plots.

2.31
Comment: Discussion: MAJOR: Where does the score threshold of 0.85 come from? Can you report the distribution of CNN output probability scores for all label categories as supplementary material? A histogram, for example, of probability scores for each label category would be interesting to see and could enhance the reader’s understanding of the challenges in labeling category 1 and 2 sounds.

Answer: The threshold of 0.85 was chosen as the cut-off point as it correlates well with the cough second measure while minimizing false-positives, results from the early ROC curves have been published (REF 13 in the manuscript).

Actions: A separate manuscript presenting a broad range of thresholds and their impact on correlation with cough seconds and true cough counts is in preparation.

2.32
Comment: Discussion: MINOR: The discussion on inter-participant variability and the debate on using solicitated voluntary coughs as a diagnostic tool is quite interesting. Does this suggest these are most likely voluntary throat clears rather than involuntary coughs? As a result, why then prompt the user to record a voluntary cough when first using the app?

Answer: The discussion does not discuss throat clears and our publication does not suggest that solicited coughs are not suitable for diagnostic purposes - more research is needed in this field, and that was not the scope of this study.

Actions: None taken.

2.33
Comment: Discussion: MINOR: What difference will the 6-tier SOP make for training the CNN model? I assume fixed 0.5sec segments are employed for training the model so how will the duration markers be more beneficial than just the onset markers used in the 4-tier SOP labeling system?

Answer: The labeling by segment, rather than by peak provides the CNN with a breadth of data around the expulsive phase.

Actions: None taken.

2.34
Comment: Discussion: MAJOR: The Pearson correlation is interesting, but I disagree that it should be the primary metric of performance. I think the authors should perform a linear

regression analysis also to analyze the slope of the regression line. Only using Pearson correlation does not disclose information regarding how many coughs are missed in detection. For example, you could theoretically have a Pearson coefficient of 1 but you are missing 1 cough in every 2 or 3 coughs consistently in your detection model (this would be reflected in the slope of the regression analysis). The Pearson correlation metric is one of a group of metrics that should be considered in my opinion. I would like to see linear regression analysis, absolute percentage error according to the total number of coughs per participant and a Bland Altman analysis.

Answer: We are adding "Table 3" to cover linear model parameter estimates: Pearson correlation, intercept and slope for both phones used.

Actions: "Table 3" added to show the linear analysis on model parameter estimates for both phones used.

2.35
Comment: Discussion: MAJOR: "...Hyfe's performance is adequate to be used in most clinical and research contexts." This study does not suggest this in my opinion if the data were collected in controlled environments, was not evaluated on real respiratory patients and the smartphones were not evaluated in multiple different positions/locations.

Answer: We agree, this phrase was overstated. Editing the manuscript.

Actions: Last paragraph of the "Discussion" was edited to: "Hyfe's performance is adequate to proceed to validation in clinical context".

2.36
Comment: Other Comments: MAJOR: While I understand the data were obtained in controlled environments, in my opinion, more data are required to obtain a more realistic measure of performance in order to suggest Hyfe could be used in clinical settings. This includes more data using different smartphone locations (in user's pocket, bag etc.), more realistic data with different types and levels of noise (including muffled coughs due to hand/mask covering) and I think it should be evaluated on a cough-by-cough basis rather 1 second windows as a form of cough rate.

Answer: This was not the scope of this study and we have changed the overstated "adequate in..clinical context" to "adequate to proceed to validation...". All the points mentioned in this comment are being addressed in a separate clinical validation study.

Actions: None taken.

2.37
Comment: Other Comments: MINOR: How does the app know if it is the user who coughed and not a person beside them in a crowded area? Has this been taken into account within the CNN training data? Please comment on this.

Answer: This was not evaluated in this study, as the scope was just to evaluate the

algorithm in detecting and recording solicited coughs in a controlled environment.

Actions: None taken.

2.38
Comment: Other Comments: MINOR: Does the Hyfe app still record audio when the user is on a phone call or sending voice messages?

Answer: This was not evaluated and not the scope of this study and the manuscript. Just for the interest - the version of the app tested in this study would have not recorded coughs on a call/voice message.

Actions: None taken.

2.39
Comment: Other Comments: MINOR: Does the app work without internet connectivity?

Answer: The connectivity details were not in the scope of this study, therefore, no information provided in the manuscript. Just for the interest - the version of the app tested in this study works without internet connectivity to capture the sounds, but not analyze them (internet connectivity was needed). The most recent version works fully on-device, with no connectivity needed to both capture and analyze the sounds.

Actions: None taken.

2.40
Comment: Other Comments: MINOR: SOP4 Supplementary Document: Page 4, Section 3, bullet 3 – should that read as "...at the end of the expulsive phase.."?

Answer: This is correct, a spelling mistake was made in "expulsive".

Actions: Correcting the spelling mistake.

2.41
Comment: Other Comments: MINOR: SOP4 Supplementary Document: I think reporting a cough as having one sound or two sounds is a little confusing. Consider moving Fig. 5 and the paragraph above it to before Fig. 3. Start with showing the sound, explaining the three phases of the cough sound event and then describe in more detail the physiology behind it. Fig. 3 panel B needs to highlight that the voiced phase is missing. It is a little confusing to just say A has a "double" sound and B has a "single sound".

Answer: We appreciate your feedback. With the SOP being updated, confusing parts have been addressed in newer editions.

Actions: None taken.

2.42

Comment: Other Comments: MAJOR: SOP4 Supplementary Document: Page 8, Table 1 – according to the 4-tier process, the label 2 may be due to the phone located in the user's pocket which, I assume, will be a vastly common occurrence if this app were to be used in cough clinical applications. In this study label 2 coughs weren't considered in the app evaluation. Why not? I think it is essential particularly for an app-based cough monitoring system to be evaluated on recordings when the phone is in the user's pocket or bag for example.

Answer: The scope of this research was not to evaluate the app in real-world environments, just the performance of the algorithm with solicited sounds in controlled environments. Further studies and validation of the system address these concerns mentioned in your comment.

Actions: None taken.

2.43
Comment: Other Comments: MINOR: SOP4 Supplementary Document: Page 8, Labeling Tips – I think the label 2 may highlight a potential limitation of the Hyfe system. It is indicating to the reader that the smartphone needs to be in close proximity to the user. Coughs may sound "distant" even if the user is coughing. Please elaborate what "distant" means here. Is this related to the audio amplitude of the cough sound? Certain types of coughs may have lower amplitudes, particularly if a very ill patient has exhausted their respiratory strength.

Answer: "Distant" in the 4-tier SOP meant subjectively distant sound (compared to surrounding sound levels both audibly and visually in the recording that is being analysed). This study was not meant to evaluate the system in clinical settings, we will be reporting on system performance with low amplitude coughs in the upcoming publications from our validation studies.

Actions: None taken.

2.44
Comment: Other Comments: MAJOR: The smartphone can be located on a table with many other people in the user's vicinity. How do you know it is the user's cough? This could be a major limitation. Particularly if the intention is to detect subtle changes in cough rate. There are many external factors when using an app-based audio cough detection system that can affect cough rate estimation. I don't think the authors make it clear how they will evaluate this or can overcome such challenges using this system.

Answer: Differentiating a cougher from the next one was not the scope of this study - this study was to evaluate the performance of the algorithm in very controlled environments - when a script is followed and the sound-producing person is known. In our validation studies and publications concerning real-world environments, all details will be provided.

Actions: None taken.

2.45
Comment: Other Comments: MINOR: If possible, it would be very beneficial to have examples of the sounds that were recorded in the controlled environment for readers to listen to.

Answer: You are requesting the sharing of recordings from participants in a trial. Following good practices, the informed consent form states that "only the study personnel will have access to the recordings". This will not be possible.

Actions: None taken.

2.46
Comment: Is the work clearly and accurately presented and does it cite the current literature? No

Answer: It would be important for the reviewer to back this assessment with specific examples of where the literature was not cited or done so inappropriately.

2.47
Comment: Is the study design appropriate and is the work technically sound? No

Answer: The vast majority of the comments by this reviewer were for broader scope than the primary objective of the study described in the manuscript.

2.48
Comment: Are sufficient details of methods and analysis provided to allow replication by others? No

Answer: The Methodology section is now updated and the full dataset and code are publicly available.

2.49
Comment: If applicable, is the statistical analysis and its interpretation appropriate? Partly

Answer: Many of the comments by this reviewer were for broader scope than the primary objective of this study described in the manuscript.

2.50
Comment: Are all the source data underlying the results available to ensure full reproducibility? No

Answer: The full data has been made available with the original version.

2.51
Comment: Are the conclusions drawn adequately supported by the results? Partly

Answer: Many of the comments by this reviewer were for broader scope than the primary

objective of this study described in the manuscript.

2.52
Comment: Competing Interests. No competing interests were disclosed by the reviewer.

Answer: This is outrageous and akin to reviewer misconduct as stated above. Please provide a full disclosure of your economic CoI or recuse.

Competing Interests: MG, EK, GG, MR and PM are employees of Hyfe Inc. Hyfe had no role in the decision to submit this protocol for publication. CCH has received consultancy fees and owns equity from Hyfe Inc. No competing interests were disclosed for all other authors.

Reviewer Report 16 August 2022

https://doi.org/10.5256/f1000research.134609.r146174

?

Vasileios Papapanagiotou
[1] Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2] Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, Thessaloniki, Greece

The paper presents an evaluation study for Hyfe. Hyfe is a smartphone app that captures audio and automatically detects coughs using machine learning methods. In this work, authors captured audio using 3 sources (an mp3 recorder for ground truth and two phones running Hyfe). Then, 3 annotators created ground truth annotations, against which Hyfe detections were compared.

The dataset includes 10-minute recordings were multiple sounds were performed in randomized sequences. The recordings were performed in complete silence. Evaluation results show high effectiveness, in particular 91% sensitivity and 98% specificity on 1,679 1-second windows.

Questions:

- In section “Methods”, subsection “Automated cough detection system”: it is mentioned that 0.5 second sound snippets are extracted and analyzed. Why was this not followed for the study and 1-second windows were used instead?

- In section “Methods”, subsection “study design” (last paragraph): how exactly were the coughs annotated? Is it with start & stop timestamp per cough? If so, please state clearly. Figure 5 is helpful and perhaps should be presented much earlier in the manuscript. Also, did each of the 3 medically trained researchers perform this process? Also, what is a “sound”

in this paragraph?

- Inter-annotator agreement would be very interesting to know.

- In section “Methods”, subsection “Data processing and analysis”: what does the “within two seconds of each other” mean? Was the maximum allowed offset between the two recordings limited by 2 seconds? If so, why? Please clarify.

- In section “Methods”, subsection “Data processing and analysis”: authors mention “if multiple explosive sounds occurred within a cough-second”. Given that Hyfe uses 0.5 second snippets and authors used 1-second windows, exactly 2 Hyfe predictions should always “fall” within one 1-second window. If so, why do the authors phrase this this way?

- Section “Methods”, subsection “Data processing and analysis”: was the 6-tier SOP used after all? If not, why is it mentioned? (mentioning it while it's not being used in the work can be confusing to the reader).

- Section “Methods”, subsection “Data processing and analysis” (last paragraph): authors mention that cough rates are of high interest in such applications. However, the data collection protocol forces some strict limitations, e.g., a 5-second silence between each activity.

- In section “Methods”, subsection “Sample size”: authors mention that “at least 385 sounds...” What is a sound in this context? Is it a 1-second window?

- In section “Results”, authors mention that in total 37 sessions were used, each being a 10-minute recording. This corresponds to 370 minutes, i.e., 22,200 seconds. However, authors also mention that the 37 sessions resulted in 672 + 1,007 = 1,679 seconds were used in the evaluation. It is not clear how the discrepancy from 22,200 to 1,679 seconds happens.

- Figure 2: what is the unit and scale for the x-axis (for both subfigures)?

- Figure 3: image quality should be improved.

- Figure 4: Authors mention that 5 seconds of silence were required (at least) between each activity. Given at least 1 second for cough, this yields a pattern of 6 seconds, and this yields a maximum of 10 coughs per minute. If so, how are values larger than 10 obtained in this plot?

- Discussion: continuous 24-hour audio recording can have a significant effect on battery consumption, this is an important limitation of the method. Also, the audio caused by the user when holding and using the phone (while it is recording audio) is also critical in evaluation of the Hyfe effectiveness.

- Discussion: "we believe that labeling cough duration rather than just its beginning has more value in further training Hyfe’s AI model". Since there is no description of the Hyfe's algorithm in the paper or any relevant experiment, this argument could be debated.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: digital signal processing, machine learning, wearables, eating behaviour

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Apr 2023

Mindaugas Galvosas

1.1
Comment: In section "Methods", subsection "Automated cough detection system": it is mentioned that 0.5 second sound snippets are extracted and analyzed. Why was this not followed for the study and 1-second windows were used instead?

Answer: Thank you for taking the time to review this paper and provide comments. Our algorithm at the time of this study was trained on a database of more than 200M of 0.5s sound snippets that are cough and cough-like sounds, collected in real life environments. Therefore, this performance evaluation also evaluated 0.5s sound snippets. Continuing our work, we are exploring the capture and analysis of 1s sound snippets, however, it was not the scope of the work described in this manuscript.

Actions: None taken.

1.2
Comment: In section "Methods", subsection "study design" (last paragraph): how exactly were the coughs annotated? Is it with start & stop timestamp per cough? If so, please state

clearly. Figure 5 is helpful and perhaps should be presented much earlier in the manuscript. Also, did each of the 3 medically trained researchers perform this process? Also, what is a "sound" in this paragraph?

Answer: In the performance evaluation study described in this manuscript, coughs were annotated following the 4-tier SOP – which instructed to indicate just the beginning of the cough sound. Further work led to developing a 6-tier and most recently – a multi-tier SOP which instructs cough and other respiratory sound annotation from the beginning to the end, taking both audio and visual (spectrogram) inputs to determine those marks.
For this study – each of the 3 medically trained researchers annotated every cough and cough-like sound following the 4-tier SOP.

Actions: Added a linear analysis (Figure 6) plot and a percentage error Bland-Altman plot (Figure 7) to the section "Results".

1.3
Comment: Inter-annotator agreement would be very interesting to know.

Answer: Thank you. We agree this is an interesting point to look at. There is a separate full paper in preparation looking specifically at this with a whole different dataset.

Actions: None taken.

1.4
Comment: In section "Methods", subsection "Data processing and analysis": what does the "within two seconds of each other" mean? Was the maximum allowed offset between the two recordings limited by 2 seconds? If so, why? Please clarify.

Answer: The manual offset was made to align the data annotation by different labelers (in an mp3 audio recording) and Hyfe timestamps on the smartphones. Due to the automated script setting, which had five second gaps between every solicited sound, two second differences were not outside the five second gap.

Actions: Added "(as this was within the silent time of five seconds between the solicited sounds in the automated script)" in the "Data processing and analysis" subsection to clarify the manual offset of the two-second time frame.

1.5
Comment: In section "Methods", subsection "Data processing and analysis": authors mention "if multiple explosive sounds occurred within a cough-second". Given that Hyfe uses 0.5 second snippets and authors used 1-second windows, exactly 2 Hyfe predictions should always "fall" within one 1-second window. If so, why do the authors phrase this this way

Answer: 2 Hyfe predictions will always be found within 1 second window but 'multiple explosive sounds' refers to a prediction of a 0.5sec snippet that contains at least 1 cough, which doesn't happen every time. Therefore, the phrasing is appropriate.

Actions: None taken.

1.6
Comment: Section "Methods", subsection "Data processing and analysis": was the 6-tier SOP used after all? If not, why is it mentioned? (mentioning it while it's not being used in the work can be confusing to the reader).

Answer: The text states "This analysis further informed the SOP… leading to the most recent version – the 6-tier SOP…" and we believe there should be no confusion here.

Actions: None taken.

1.7
Comment: Section "Methods", subsection "Data processing and analysis" (last paragraph): authors mention that cough rates are of high interest in such applications. However, the data collection protocol forces some strict limitations, e.g., a 5-second silence between each activity.

Answer: While rates indeed remain interesting, calculating them was outside the scope of this performance evaluation, nor is it possible with the methodology that was followed. The main purpose of this work was to evaluate the performance of the system capturing and recording solicited coughs and cough-like sounds in controlled environments.

Actions: None taken.

1.8
Comment: In section "Methods", subsection "Sample size": authors mention that "at least 385 sounds…" What is a sound in this context? Is it a 1-second window?

Answer: Sound refers to any cough or other sound produced by the study participants according to the script provided.

Actions: None taken.

1.9
Comment: In section "Results", authors mention that in total 37 sessions were used, each being a 10-minute recording. This corresponds to 370 minutes, i.e., 22,200 seconds. However, authors also mention that the 37 sessions resulted in 672 + 1,007 = 1,679 seconds were used in the evaluation. It is not clear how the discrepancy from 22,200 to 1,679 seconds happens.

Answer: 1,679 was the number of solicited sounds that were evaluated, not the total number of seconds of recording for all participants. Further explanation is provided in the "Study design" section.

Actions: Improved "Study design" section: "Participants were asked to produce a series of

solicited sounds by reading a provided script, while being recorded with an MP3 recorder and monitored by Hyfe on two identical smartphones. The phones and recorder were placed on a table at approximately 50 cm from the participants, with microphones oriented towards them. A pre-generated computer script instructed participants to produce a series of 46 sounds, of which 18 were coughs, the rest consisted of solicited sneezes, throat clearings, spoken letters or words in the same 10 minutes. Participants were instructed to cough once every time they were prompted by the script to do so. In total for each participant, the script included instructions to cough 20 (18 as individual coughs and 2 in the literary text) times, sneeze 10 times, clear throat 5 times and produce 15 sounds (explosive words, for example, “paella” and numbers as “93”). “

1.10
Comment: Figure 2: what is the unit and scale for the x-axis (for both subfigures)?

Answer: The name of the x-axis in this version is called “frequency of observation”, however, after another review with our data science team, we have agreed to change this Figure 2 to a new visual - newly submitted histograms in Figure 2, which convey the same information on specificity and sensitivity for both phones more clearly and, in our opinion, more efficiently.

Actions: “Figure 2” updated to a new visual. The “Figure 2” is also now accompanied by the “Table 2: Summary statistics on sensitivity and specificity for both phones used”.

1.11
Comment: Figure 3: image quality should be improved.

Answer: Thank you for noting, the image will be replaced with a better quality file.

Actions: New image generated and uploaded (Figure 3).

1.12
Comment: Figure 4: Authors mention that 5 seconds of silence were required (at least) between each activity. Given at least 1 second for cough, this yields a pattern of 6 seconds, and this yields a maximum of 10 coughs per minute. If so, how are values larger than 10 obtained in this plot?

Answer: Some participants were instructed to cough once, however, they produced two coughs. These coughs were annotated by the labelers and also picked up by Hyfe, when they happened with at least 0.5s of separation.

Actions: None taken.

1.13
Comment: Discussion: continuous 24-hour audio recording can have a significant effect on battery consumption, this is an important limitation of the method. Also, the audio caused by the user when holding and using the phone (while it is recording audio) is also critical in evaluation of the Hyfe effectiveness.

Answer: This is a valid point; however, this performance evaluation was done in laboratory settings and was not meant to evaluate battery usage and performance. Additionally, directionality of the audio source was not evaluated by this study and is being evaluated with the continued studies.

Actions: None taken.

1.14
Comment: Discussion: "we believe that labeling cough duration rather than just its beginning has more value in further training Hyfe's AI model". Since there is no description of the Hyfe's algorithm in the paper or any relevant experiment, this argument could be debated.

Answer: For model training purposes we have seen that labeling the full duration of the sound is a better way of annotating data, ensuring that the algorithm is being trained on the full duration of cough sound. As convolutional neural networks are employed by Hyfe - they learn by example. As long as the training data is relatively unbiased and representative, a neural net can identify a "feature" (such as the acoustic signature of a cough) in a myriad of samples, even if those samples do not resemble each other.

Actions: Added to the "Discussion": "As convolutional neural networks are employed by Hyfe - they learn by example. As long as the training data is relatively unbiased and representative, a neural net can identify a "feature" (such as the acoustic signature of a cough) in a myriad of samples, even if those samples do not resemble each other."

1.15
Comment: Are all the source data underlying the results available to ensure full reproducibility? – replied "No".

Answer: All source data is updated and made available to reproduce all the analysis presented in the manuscript.

Actions: None taken.

1.16
Comment: Are the conclusions drawn adequately supported by the results? – replied "Partly"

Answer: We believe that the proposed changes will strengthen the conclusions drawn from the results, thank you for your input and feedback.

Actions: None taken.

Competing Interests: MG, EK, GG, MR and PM are employees of Hyfe Inc. Hyfe had no role in the decision to submit this protocol for publication. CCH has received consultancy fees

and owns equity from Hyfe Inc. No competing interests were disclosed for all other authors.