BMJ Open Respiratory Research

Performance evaluation of human cough annotators: optimal metrics and sex differences

Isabel Sanchez-Olivieri,[1] Matthew Rudd,[2] Juan Carlos Gabaldon-Figueira,[3] Francisco Carmona-Torre,[1] Jose Luis Del Pozo,[1] Reid Moorsmith,[2] Lola Jover,[2] Mindaugas Galvosas ,[2] Peter Small,[2] Simon Grandjean Lapierre,[4,5] Carlos Chaccour [1,3,6]

To cite: Sanchez-Olivieri I, Rudd M, Gabaldon-Figueira JC, *et al*. Performance evaluation of human cough annotators: optimal metrics and sex differences. *BMJ Open Respir Res* 2023;**10**:e001942. doi:10.1136/bmjresp-2023-001942

SGL and CC contributed equally.

Received 8 July 2023
Accepted 31 October 2023

[1]Universidad de Navarra, Pamplona, Spain
[2]Hyfe Inc, Wilmington, Delaware, USA
[3]ISGlobal, Barcelona institute for Global Health, Barcelona, Spain
[4]Dept of Microbiology, Infectious Diseases and Immunology, Research Center of the University of Montreal Hospital Center, Montreal, Quebec, Canada
[5]Immunopathology Axis, Research Center of the University of Montreal Hospital Center, Montreal, Quebec, Canada
[6]Centro de investigación biomédica en red enfermedades infecciosas, Madrid, Spain

Correspondence to
Dr Carlos Chaccour;
carlos.chaccour@isglobal.org

ABSTRACT

Introduction Despite its high prevalence and significance, there is still no widely available method to quantify cough. In order to demonstrate agreement with the current gold standard of human annotation, emerging automated techniques require a robust, reproducible approach to annotation. We describe the extent to which a human annotator of cough sounds (a) agrees with herself (intralabeller or intrarater agreement) and (b) agrees with other independent labellers (interlabeller or inter-rater agreement); we go on to describe significant sex differences in cough sound length and epochs size.

Materials and methods 24 participants wore an audiorecording smartwatch to capture 6–24 hours of continuous audio. A randomly selected sample of the whole audio was labelled twice by an expert annotator and a third time by six trained annotators. We collected 400 hours of audio and analysed 40 hours. The cough counts as well as cough seconds (any 1 s of time containing at least one cough) from different annotators were compared and summary statistics from linear and Bland-Altman analyses were used to quantify intraobserver and interobserver agreement.

Results There was excellent intralabeller (less than two disagreements per hour monitored, Pearson's correlation 0.98) and interlabeller agreement (Pearson's correlation 0.96), using cough seconds as the unit of analysis decreased annotator discrepancies by 50% in comparison to coughs. Within this data set, it was observed that the length of cough sounds and epoch size (number of coughs per bout or attach) differed between women and men.

Conclusion Given the decreased interobserver variability in annotation when using cough seconds (vs just coughs) we propose their use for manually annotating cough when assessing of the performance of automatic cough monitoring systems. The differences in cough sound length and epochs size may have important implications for equality in the development of cough monitoring tools.

Trial registration number NCT05042063.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Human annotation is the gold standard for quantitative assessment of cough. The agreement between human labellers will vary according to the unit of analysis used. Cough seconds (any second of time containing at least one cough) are highly correlated with actual cough numbers.

WHAT THIS STUDY ADDS

⇒ We use a large cough-dataset sequentially labelled by different annotators and describe the variation in agreement according to the unit of analysis chosen. Additionally, we describe differences in cough-sound length and cough epochs that are attributable to the patient's sex.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The higher reproducibility in the labelling of cough seconds (vs just coughs) imply this unit of analysis can be used for validation of emerging automatic cough-counting devices. The sex differences in cough length can have implications for disease transmission, diagnosis and health-seeking behaviour.

INTRODUCTION

Cough is a key symptom of most respiratory diseases and is among the most frequent reasons for seeking medical attention.[1] However, outside of brief interactions during a clinic visit, healthcare providers have little quantitative insight into a patient's cough and must rely on patient-reported outcomes which are subject to recall and other forms of bias.[2 3] Quantifying cough for 24 hours is now possible with several semiautomated or fully automated systems.[4 5] However, the specific methods they use for annotating cough are not publicly described with enough detail to be reproduced. Furthermore, given the stochastic nature of cough patterns, even 24 hours of monitoring time can lead to the mistaken conclusions about a patient's health status or the effectiveness of prospect antitussive drugs.[6 7] The emergence of machine

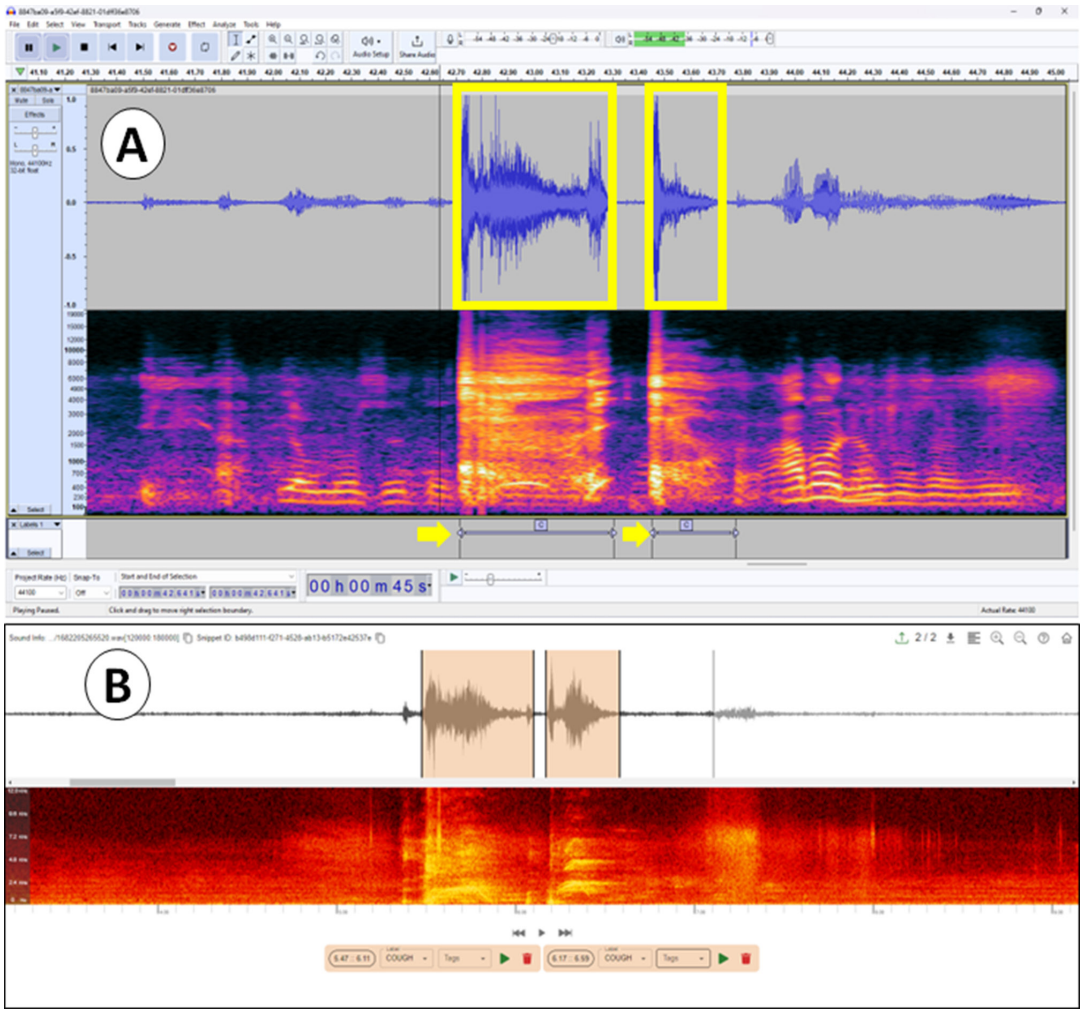

Figure 1 Labelling of two contiguous coughs in (A) Audacity, yellow boxes added for clarity on start and end of cough-segment labels which are below marked by yellow arrows and (B) Hyfe's browser app.

Table 1 Participant characteristics

Female, n (%)	13 (54)
Inpatients, n (%)	16 (66)
Age, mean (range)	63 years (29–91)
Hours monitored, mean (range)	16 hours 46 min (5:05–24:51)
Hourly cough rate, mean (range)	11.8 (0.5–35.2)
Diagnoses, n (%)	
COVID-19*	14 (58)
Bacterial pneumonia	3 (12)
Upper respiratory tract infection	2 (8)
Asthma exacerbation	2 (8)
Acute bronchitis	1 (4)
COPD exacerbation	1 (4)
Dyspnoea	1 (4)

*With/without associated bacterial pneumonia.

learning now allows for continuous, unobtrusive cough monitoring for extended periods of time.[6 8]

The Hyfe cough monitoring system is one of such emerging tools. It leverages artificial intelligence on a smartphone or smartwatch platform to automatically and unobtrusively detect and quantify cough in varied environments and real-world acoustic conditions. The clinical validation of this and other new systems for regulatory purposes will require a robust gold standard to serve as comparator. As human annotation remains the most frequently used comparator for cough detection systems,[9 10] an in-depth understanding of the operator-dependent variations in cough metrics is required.

Here, we describe the process of human cough labelling followed at Hyfe, its intra-annotator and interannotator agreement, and the user experience of two available cough-labelling software products. Additionally, we describe a novel finding regarding differences in the length of cough-sound and number of coughs per epoch in men and women in this cohort of patients.

Table 2 Total and percentage labels by category, labeller and round

Label category	Expert labeller first pass	Expert labeller second pass	Six labellers third pass
Cough	803 (52%)	834 (49%)	500 (57.6%)
Cough—far	46 (2.9%)	70 (4.1%)	18 (2%)
Throat clear	237 (25.7%)	405 (23.8%)	193 (22.2%)
Throat clear—far	208 (13.4%)	318 (18.7%)	16 (1.8%)
Sneeze	2 (0.1%)	1 (0.06%)	2 (0.2%)
Sneeze—far	0	0	0
Other	55 (3.5%)	46 (2.7%)	39 (4.4%)
Other—far	33 (2.1%)	26 (1.5%)	1 (0.1%)
Total	1544	1700	868

MATERIALS AND METHODS

Study subjects

Outpatients and inpatients older than 18 years presenting with a main complaint of cough to the Clínica Universidad de Navarra, Pamplona, Spain between November 2021 and May 2022 were invited to participate.

Data collection

Basic demographic, clinical data and diagnosis were collected from all participants at enrolment. Participants were instructed to simultaneously use a dedicated Android smartphone (Motorola G30) running Hyfe Cough Tracker and an active MP3 recorder (Sony ICD-PX470). Both devices were carried in a shoulder bag during daytime or were placed on top of the bedside table during sleeping hours. Both devices were used continuously for a minimum of 6 hours and a maximum of 24 hours per participant.

Description of the tool

Hyfe is an AI-enabled mobile phone app that detects and captures short snippets (0.5 s) of explosive (peak) sounds and then classifies them as cough or non-cough using a convolutional neural network model,[11 12] data processing is done on device which offers a robust privacy protection. Previous assessments of its performance in controlled and real-world settings show high reliability as well as correlation with clinical changes and treatment response.[6 11 13–15]

Acoustic data labelling

Continuous audio from the audiorecordings was manually reviewed and annotated by trained labellers following a pre-established standard operating procedure.[15] In brief, segment-length labels were placed over the sounds of interest starting at the point where acoustic background was modified (for coughs this is the start of the explosive phase) through the moment when the acoustic background returned to normal (for coughs this is usually at the end of the vocal phase).[16] This cough labelling method produces two timestamps for each cough, one at the beginning and one at the end of each sound of interest. Epochs are defined as several explosive phases with less than 2 s between them. When these occurred, each cough (explosive and vocal phases) was labelled individually. Peak sounds were marked as coughs, throat clears, sneezes or 'other.' This last category was applied only to loud, potentially cough-like peaks according to the annotator. If the labeller subjectively perceived the sounds as faint or occurring far away from the recorder, the sublabel 'far' could be added, this was done to assess the potential acoustic contamination by non-participant coughs.

All annotations were done in duplicate, using the freely available Audacity software (Audacity team (2021). Audacity(R): Free Audio Editor and Recorder (Computer application) V.3.1.3) as well as a browser-based app developed by Hyfe (https://hyfe-continuous-labeling.web.app) (figure 1).

Table 3 Intraobserver agreement

	Linear analysis			Bland-Altman analysis			
Unit	Correlation	Slope	Intercept	Bias*	MOE*	Slope	Intercept
Coughs	0.9841	1.0502	−0.0061	0.0789	1.5137	0.0654	−0.0346
Fixed cough seconds	0.9915	1.0182	0.0052	0.0299	0.8350	0.0267	−0.0067
Mobile cough seconds	0.9938	1.0108	0.0092	0.0213	0.5976	0.0170	0.0020

*Average of the differences between the paired cough counts; the margin of error is twice the SD of these differences.
MOE, margin of error.

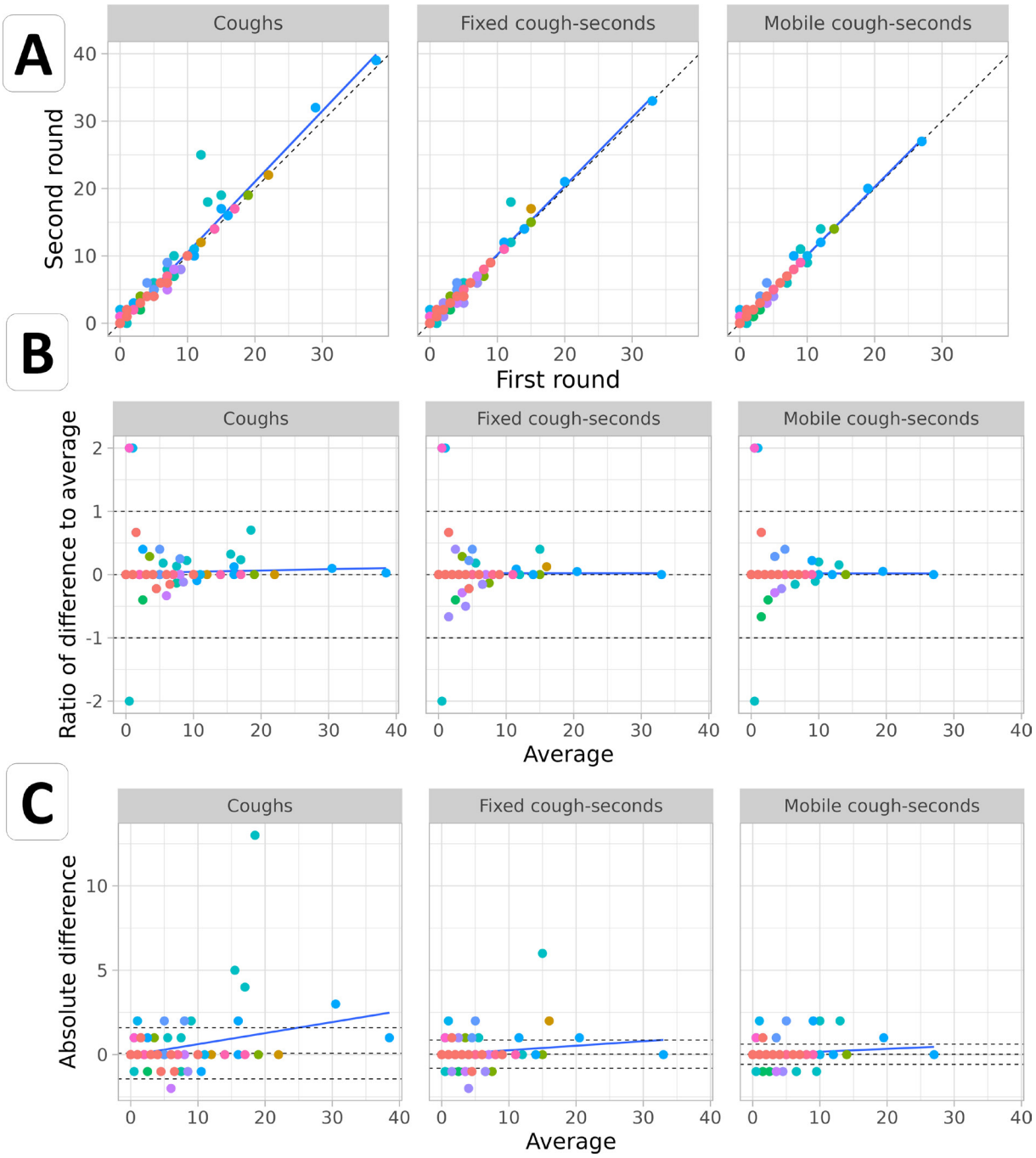

Figure 2 Intraobserver agreement. Linear analysis; each dot represents one person-hour, different colours represent different labellers, dashed line is the line of perfection, blue line is the best fit (A). Intraobserver agreement. Bland-Altman analysis for absolute difference (B). Intraobserver agreement. Bland-Altman analysis for ratio of difference to average (C).

Labellers were blinded to their previous work as well to one another's labels. The audio of monitoring sessions was divided into 5 min files. A randomly selected 10% of these 5 min files, totalling 40 hours across all participants, was selected for labelling.

Table 4 Intralabeller agreement by unit of analysis

	Consensus	Disagreements	Disagreements per hour
Coughs	775	77	1.97
Fixed cough seconds	618	50	1.27
Mobile cough seconds	514	40	1.02

Table 5 Interobserver agreement

	Linear analysis			Bland-Altman analysis			
Unit	Correlation	Slope	Intercept	Bias	MOE	Slope	Intercept
Coughs	0.9629	0.8643	0.1047	−0.2287	2.6689	−0.1099	0.0289
Fixed cough seconds	0.9676	0.8753	0.0757	−0.1659	1.9767	−0.1019	0.0229
Mobile cough seconds	0.9736	0.8946	0.0536	−0.1166	1.5125	−0.0858	0.0169

MOE, margin of error.

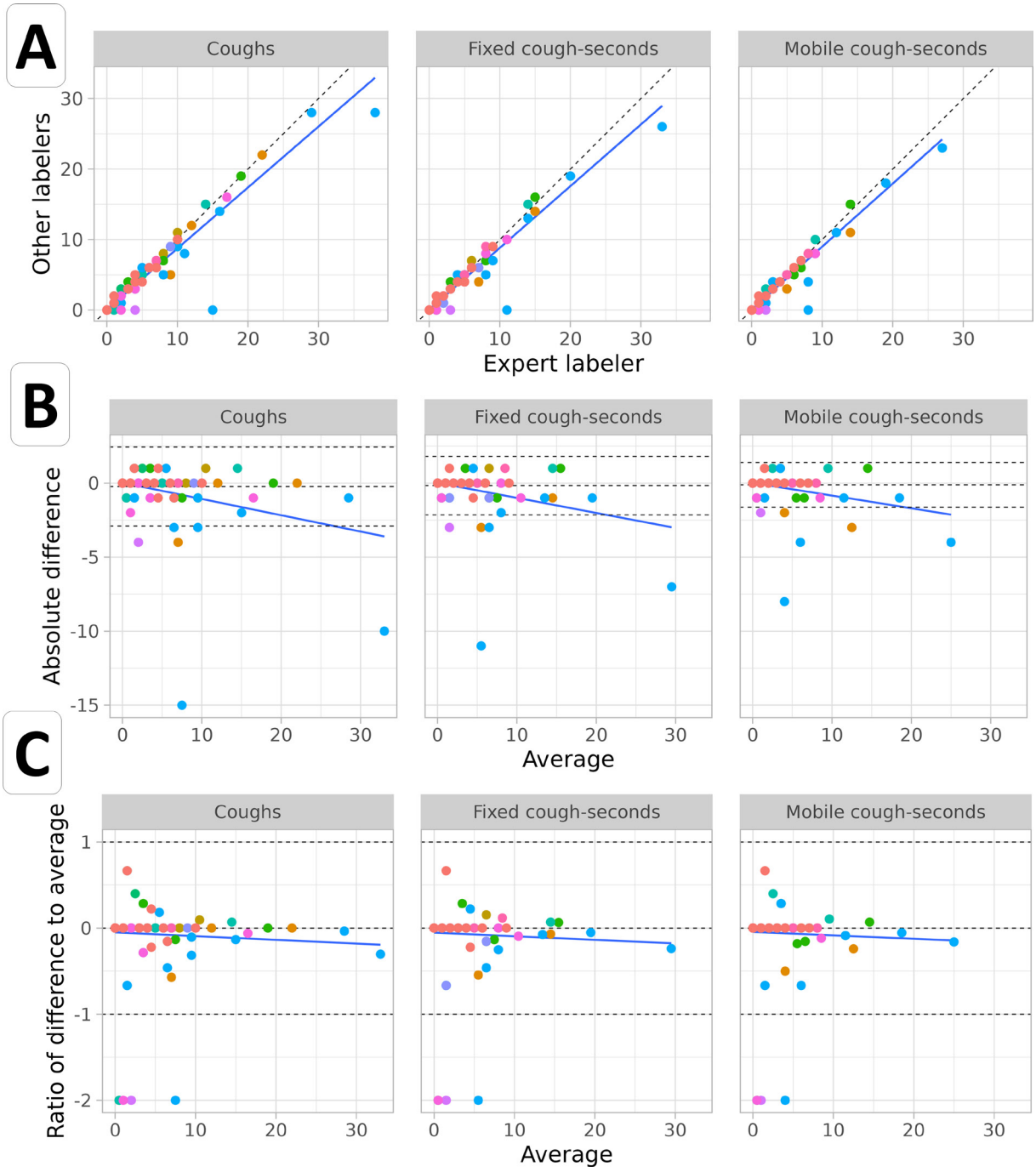

Figure 3 Interobserver agreement. Linear analysis; each dot represents one person-hour, different colours represent different labellers, dashed line is the line of perfection, blue line is the best fit (A). Interobserver agreement. Bland-Altman analysis for absolute difference (B). Interobserver agreement. Bland-Altman analysis for ratio of difference to average (C).

Table 6 Unit agreement statistics

Comparison	Round	Correlation	Slope	Intercept
Fixed cough seconds versus coughs	Round 1	0.9862	0.7883	0.0199
Fixed cough seconds versus coughs	Round 2	0.9864	0.7588	0.0398
Mobile cough seconds versus coughs	Round 1	0.9812	0.6669	−0.0025
Mobile cough seconds versus coughs	Round 2	0.9824	0.6364	0.0203
Mobile versus fixed cough seconds	Round 1	0.9878	0.8400	−0.0112
Mobile versus fixed cough seconds	Round 2	0.9887	0.8327	−0.0048

To assess the intralabeller agreement, an expert annotator with over 800 hours of labelling experience labelled each selected 5 min file twice. To assess the interlabeller agreement, a third review was conducted by a group of six other annotators (MG, PS, RM, LJ and CC) labelled a subsample of 20 hours of audio a third time.

Units of analysis

Two basic units of analysis were used:

a. Coughs, as a segment individually time-stamped by human annotators as described above.
b. Cough seconds, derived automatically from (a), defined as any 1 s time span containing at least one cough (as defined in (a)).

Cough seconds are a valuable alternative measure of accuracy given that coughs can occur in rapid succession, with or without intervening inhalation. In patients with multiple epochs, it becomes difficult, even for trained annotators, to distinguish between the end of one cough and the beginning of the next cough.

For cough seconds, we used two approaches to define their starting time:

a. Fixed cough seconds, with a start time determined by rounding a cough's timestamp down to the nearest preceding clock second.
b. Mobile cough seconds, with starting time being the precise timestamp (to millisecond precision) of the first cough among a group of coughs occurring within the subsequent second.

For cough metrics, only the peak sounds labelled as 'coughs' were used. Peak sounds containing the sub label 'far' were considered non-coughs and excluded from analysis.

Analysis

The cough counts from the same or different annotators were compared and summary statistics from linear and Bland-Altman analyses were used to quantify intraobserver and interobserver agreement. For each

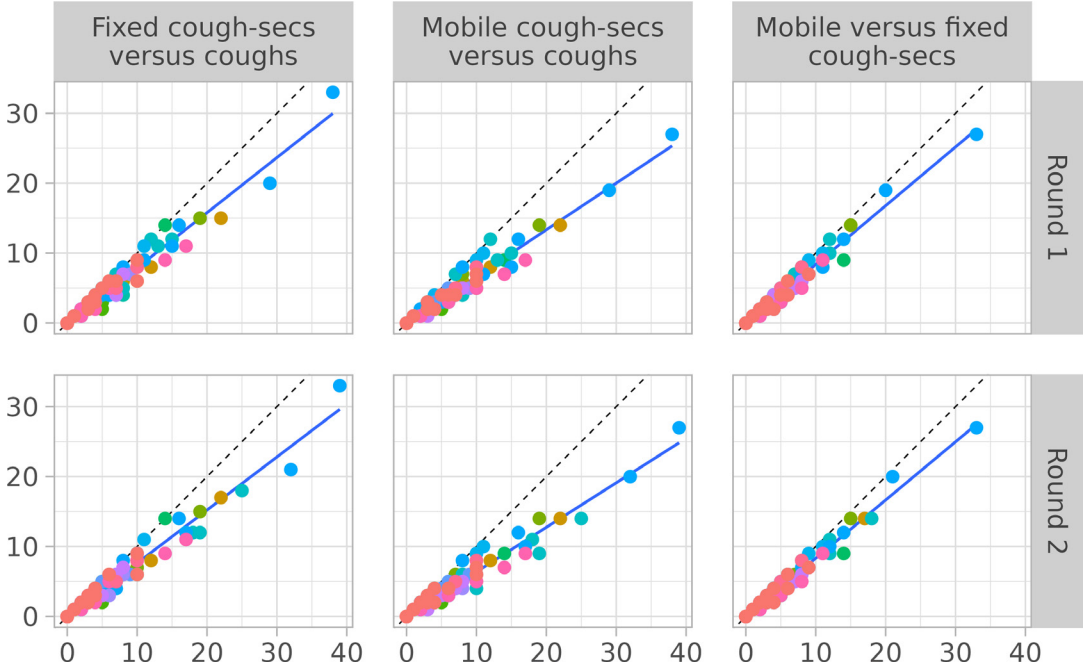

Figure 4 Agreement between labelling rounds and unit of analysis. Each dot represents one person-hour, different colours represent different labellers, dashed line is the line of perfection, blue line is the best fit.

Table 7 Summary statistics of cough length data

Analyses	Criteria	Sample	Q1	Median	Mean	Q3	IQR	P value*
Overall	Exp. labeller	803	0.32	0.40	0.45	0.53	0.20	0.1490
	Exp. labeller	834	0.31	0.39	0.43	0.52	0.20	
	six labellers	500	0.31	0.38	0.43	0.51	0.20	0.3637
	Overall	2137	0.31	0.39	0.44	0.52	0.20	
By sex patients (coughs)	Male	11 patients 955 coughs	0.42	0.48	0.50	0.57	0.14	0.025
	Female	12 patients 1182 coughs	0.36	0.38	0.40	0.44	0.08	
By diagnosis patients (coughs)	COVID-19	13 patients 1438 coughs	0.37	0.42	0.44	0.45	0.07	0.492
	All other	10 patients 699 coughs	0.39	0.47	0.47	0.53	0.13	

*Adjusted for clustering by patient.

analysis (intraobserver or interobserver agreement) the following metrics were calculated and presented in tables:

- Pearson correlation coefficient: quantifies the strength of the linear association.
- Bias: the average difference between paired cough counts.
- Bias margin of error: twice the SD of the differences between paired cough counts.
- Slope: the slope of the least squares line of best fit, for both paired cough counts (linear analysis) and differences versus averages (Bland-Altman analysis).
- Intercept: the intercept of the least squares line of best fit, for both paired cough counts (linear analysis) and differences versus averages (Bland-Altman analysis).

For each analysis, scatterplots and Bland-Altman plots were drawn to provide a visual summary.

Finally, to examine the relationships between different units of analysis (coughs and cough seconds), we applied a linear analysis (correlation, slope, intercept, scatterplot) to the expert annotator's two rounds of labels.

Cough sound length and patterns by sex and diagnosis

The durations of all labelled coughs were estimated using the start and end time stamps. Evidence of differences in mean cough duration by sex and by diagnosis (COVID-19 vs all other) were assessed with two-sample t-tests corrected for clustering. Each patient was a cluster with a number of coughs.

The numbers of epochs containing one cough, two coughs, three coughs and four or more coughs were calculated overall, by sex and by diagnosis. Evidence of differences in the resulting distributions of cough epoch sizes by sex and by diagnosis (COVID-19 vs all other) were assessed with χ^2 tests.

Annotator experience with two different software products

The annotators were asked to write down the advantages and disadvantages of Audacity and Hyfe's browser-based labelling app. The recurrent topics were identified by one of the researchers (CC) and tabulated descriptively.

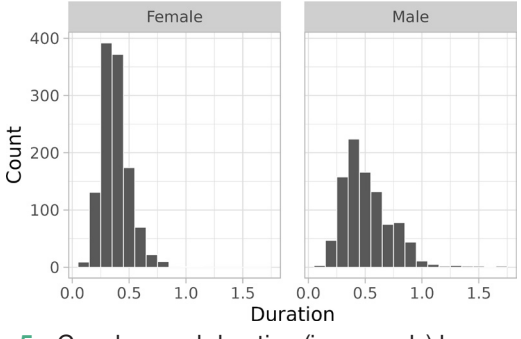

Figure 5 Cough sound duration (in seconds) by sex of 23 patients encompassing 2137 coughs (one participant did not have any cough labels in the randomly selected segments).

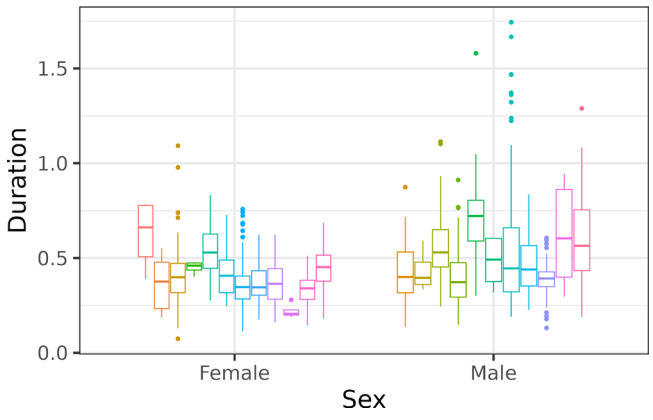

Figure 6 Cough length distribution by sex.

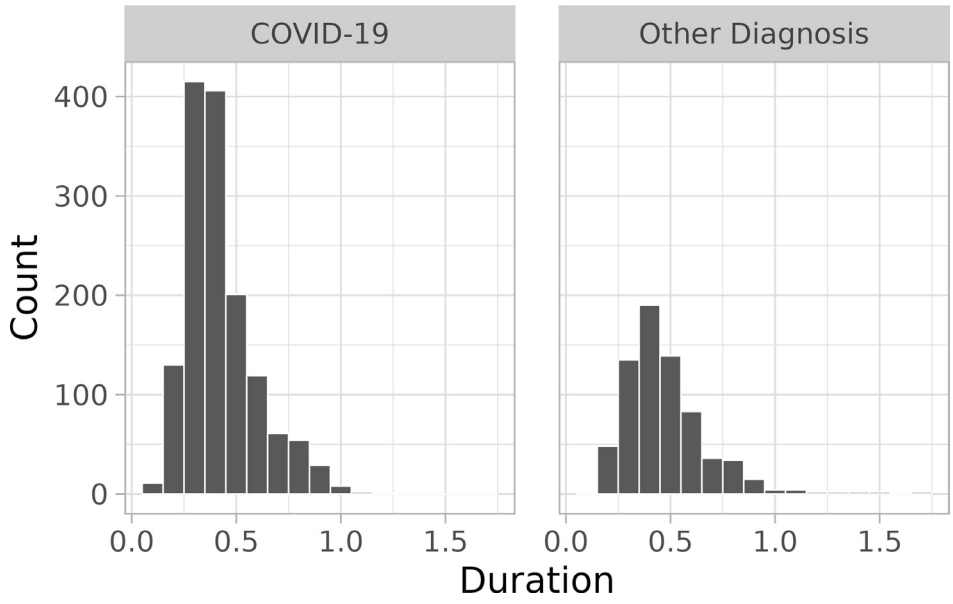

Figure 7 Cough sound duration by diagnosis 23 patients encompassing 2137 coughs (1 participant did not have any cough labels in the randomly selected segments).

Patient and public involvement

Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

RESULTS

Participants

Out of 32 participants invited, 24 consented to participate and were enrolled. All enrolled participants complied with the study monitoring requirements. Sixteen were recruited as inpatients in private rooms and eight were recruited as outpatients. The mean age was 63 (range 29–91). Thirteen were female. The tracking time ranged from 5 hours and 5 min to 25 hours and 15 min (mean 16 hours and 46 min). The mean hourly cough rate was 12 (range 0.5–35 coughs/hour). The most frequent diagnosis was COVID-19 (14/24). The demographics, diagnosis and tracking data of all participants are provided in table 1.

Total monitored time and labels

A total of 402 hours and 32 min of audio was captured from all participants. The 10% selected for labelling included 40 hours and 10 min divided into 482 files of 5 min each.

On the 40 hours and 10 min sample selected for annotation, the expert reviewer placed 1544 labels on her first pass (803 coughs) and 1700 labels on her second pass (834 coughs). The group of 6 annotators labelling 20 hours of audio placed 868 labels (500 coughs) on the third labelling pass. Close to 51% of all labels placed corresponded to coughs (2137/4112) of which only 134 (3 % of all labels) were coughs classified as 'far' by the annotators. Less than 5% of all labels corresponded to sounds other than coughs, throat clears or sneezes. The total number of labels placed by category, annotator and pass are shown in table 2. There was no difference in the hourly cough rate of inpatients and outpatients (16.3 vs 12.2, respectively, p=0.36).

Intraobserver agreement

Labels made blindly by the same listener in separate sessions had a Pearsons's correlation of 0.98 or above regarding coughs, fixed cough seconds and mobile cough seconds (table 3, figure 2).

The consensus, disagreements and disagreements per hour of the expert annotator for coughs, fixed cough seconds and mobile cough seconds are presented in table 4.

Table 8 Cough metrics by sex and disease

	COVID-19		P value	Other diagnosis		P value
	Male	Female		Male	Female	
Total coughs	6 patients 490 coughs	7 patients 948 coughs		5 patients 465 coughs	5 patients 234 coughs	
Mean cough length	0.54	0.36	0.025	0.50	0.43	0.719

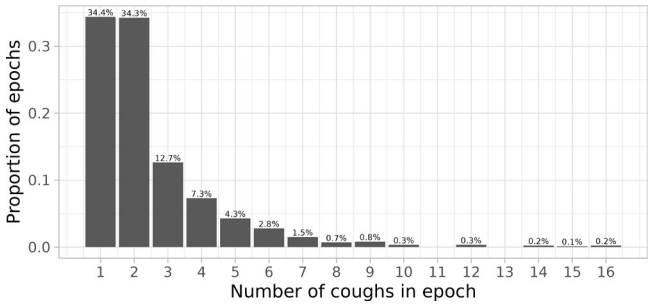

Figure 8 Histogram of cough-epoch sizes.

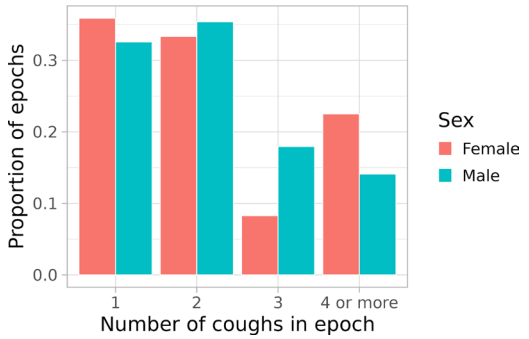

Figure 9 Distribution of cough epoch size by sex.

The intraobserver agreement for far-labelled coughs had a correlation of 0.635 and a slope of 0.98. This is similar to the results obtained including all far-labelled sounds (coughs, throat clears and others) with correlation 0.694 and slope 0.962.

Interobserver agreement

Labels placed blindly by the group of six annotators had a Pearson's correlation of 0.96 or higher for coughs, fixed cough seconds and mobile cough seconds when compared with labels of the first and second pass (table 5, figure 3).

The interobserver correlation for far-labelled coughs had a correlation of 0.506 and a slope of 0.608. The results for all far-labelled sounds show a correlation of 0.455 and a slope of 0.223.

Unit agreement statistics

The relationship between each unit of analysis and each labelling round done by the first labeller is presented in table 6 and figure 4.

Cough sound length and patterns by sex and diagnosis

Summary statistics of cough sound length data is presented in table 7. The mean duration of the 2137 cough labels placed was 0.44 s (median 0.39, IQR 0.20). There was no difference in the mean label duration placed by the expert annotator in the two rounds (mean 0.45 vs 0.43, p=0.14). Nor was there a difference between the length of the labels placed by the expert annotator and that of the six other labellers (mean 0.44 vs 0.43, p=0.36).

Of the 2137 cough labels placed, 955 (44.7%) corresponded to male participants and 1182 (55.3%) corresponded to female participants. The length of cough sounds from female participants was 20% shorter than that of male participants (mean 0.40 vs 0.50 s, p=0 0.025) (table 7 and figure 5). Additionally, men had a higher variance (figure 6).

Of the 2137 cough labels placed, 1438 (67.3%) corresponded to 13 participants with an underlying COVID-19 diagnosis and 699 (32.7%) to 10 participants with other diagnoses. There was no difference in the length of cough sounds from COVID-19 participants and that of participants with all other diagnoses (mean 0.44 vs 0.47 s, p=0.49) (table 7 and figure 7).

After stratifying by disease, the sex-related differences in cough sound length remained in those with COVID-19 but not between male and female participants with other diagnoses (table 8).

Of the 2137 cough labels placed, 296 (34.38%) corresponded to single coughs, 295 (34.26%) to epochs with two coughs, 109 (12.66%) to epochs with three coughs and 161 (18.7%) epochs with for or more coughs. The longest epochs labelled included 16 coughs and occurred twice in the labelled audio (0.23% of all labelled coughs) (figure 8).

There is a statistically significant difference in the epoch size distribution between male and female participants (table 9 and figure 9). Women had 77% of coughs in epochs of 3 or less coughs, while men had 86% of coughs in epochs of 3 or less coughs.

There is a statistically significant difference in the epoch size distribution between participants with COVID-19 and those with other diagnoses (table 10 and figure 10). Participants with COVID-19 had a higher proportion of single coughs.

Annotator experience with two different software

The summary of the annotator experience with the two software products is presented in table 11.

Table 9 Distribution of cough epoch size by sex

Epoch size	1		2		3		≥4		χ^2 P value
Sex	Female	Male	Female	Male	Female	Male	Female	Male	
n (% by sex)	169 (35.88)	127 (32.56)	157 (33.33)	138 (35.38)	39 (8.28)	70 (17.95)	106 (22.51)	55 (14.10)	0.00001738

Table 10 Distribution of cough epoch size by diagnosis

Epoch size	1		2		3		≥4		χ^2 P value
COVID-19	No	Yes	No	Yes	No	Yes	No	Yes	
n (% by Dx)	72 (25.71)	224 (38.55)	102 (36.43)	193 (33.22)	54 (19.29)	55 (9.47)	52 (18.57)	109 (18.76)	0.00002471

DISCUSSION

Here, we describe a rigorous method for manually annotating coughs from continuous audio recordings. We also describe how different units of cough analysis can be applied to these annotations and use summary statistics from linear and Bland-Altman analyses to assess the agreement within and between human annotators and using different metrics. Finally, we analyse the individual cough length and epoch size by sex and diagnosis.

The cough recordings used in this study come from a diverse group of participants which allowed for the evaluation of the cough metrics and annotation process in samples from males and females across a range of different ages and acoustic environments. COVID-19 was the most prevalent cause of cough at the time of the study, this sample was however enriched by over 40% of participants with other causes of cough.

We collected over 400 hours of continuous audio and human annotators labelled a random 10% plus the full 24 hours of two patients compatible with the intended use population of a cough monitor. From this sample of 4112 labels were placed encompassing a broad range of possible labels including coughs (over 50% of all labels), throat clears, sneezes and other cough-like sounds.

It is known that human annotators show better agreement in cough counts than in the assessment of other aspects of cough such as severity, strength or quality.[17 18] In our data, there was high intraobserver and interobserver agreement on numbers of coughs per unit of time regardless of units of measure used or analysis strategy, this is aligned with previous studies on this topic[18–23] as well as studies assessing agreement on the numerical assessment of other clinical processes such as the respiratory rate.[24]

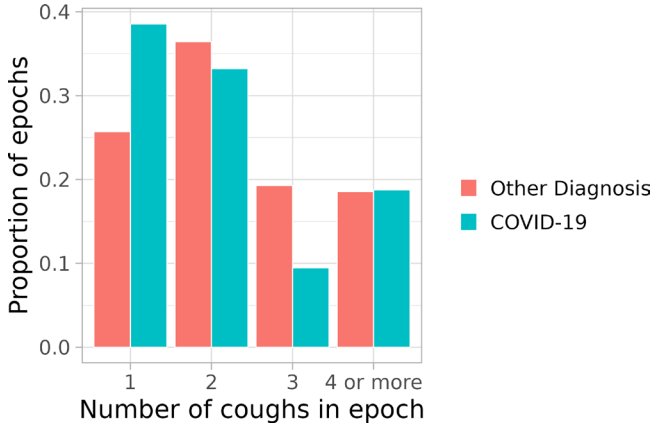

Figure 10 Distribution of cough epoch size by diagnosis.

Far-labelled coughs assessed by the same annotator show a high numerical slope with a correlation of 0.6, which suggests that numerical corrections can improve the performance, however, the low correlation obtained interlabeller means the 'far' tag will not suffice to assess acoustic contamination on its own.

We described several different ways to describe cough frequency. As previously described, coughs, mobile cough seconds and fixed cough seconds are all highly correlated and either could be used to reflect a user's cough applying a correction factor between them.[2] The number of intra-annotator disagreements per hour, although low across all cough metric units, is reduced almost 50% when mobile cough seconds are used versus just coughs (table 4). Hence, for evaluating automatic systems, mobile cough seconds are a more reproducible metric. All three metrics are highly correlated and can be reported in a way that has intuitive value to patients and providers as 'coughing rate per hour'.

This study also explored two different ways to quantitatively assess the accuracy of automated cough monitoring in comparison to human listeners. Comparing on an event-by-event basis in terms of specific coughs is problematic due to the inability of annotators to discern when sequential explosive peaks are separate coughs or part of the same cough epoch, as this report demonstrated. Using that approach to compare on the basis of specific cough seconds minimises this source of noise and allows for calculating the sensitivity and specificity for individual coughs of an automatic cough counter. This approach to describing performance is standard in many settings, such as diagnostic testing. However, performance at the level of individual coughs is not appropriate as the clinically relevant question is about coughing trends and totals, not any single specific cough, hence best expressed as a cough frequency or rate. Thus, we propose that the clinically relevant performance metric is correlation of hourly cough second rate which we can express as a Pearson correlation, y intercept and slope and are highly correlated with 'raw' cough rate.

Manual cough counts are the most commonly used gold standard to determine accuracy of automatic cough detectors.[9 10] Recorded audio has been shown equivalent to video recordings to assess cough frequency.[25] When involving human annotators, a visual depiction of the sounds has been proven useful.[23] While Audacity has been used successfully in the past, it requires manual handling

Table 11 Annotators perceptions after working with Audacity and Hyfe's browser-based app

	Advantages	Disadvantages
Audacity	► Mistakes can be corrected retrospectively ► Work on a label while listening to the audio	► Requires manual download of labels and manual upload to a database ► All actions require using the mouse ► Continuous listening to the same participant makes it easier to get distracted ► Selecting one specific label is laborious
Hyfe's browser-app	► Very intuitive ► All labels automatically uploaded at backend ► All actions can be done with the keyboard ► Can mix a patient's segments, stimulating focus ► Label selection is easy	► Mistake correction requires interaction with an administrator ► Segmentation makes it easier to have partial sounds at the end

of data which is prone to human error, Hyfe's labelling software is reported by annotators to be easy to use and allows for automatic management of databases in the backend, reducing the likelihood that errors are made in managing data. The completed and ongoing studies using here described annotation protocol has instructed the most recent update to the continuous cough sound annotation standard operating procedure, which can be accessed here.

There are limited published data on the duration of cough sounds.[16 26 27] The normal duration of cough sound varies in the literature, from 0.3 to 1.0 s[27] with lengthening described in association with disease or smoking status.[27] Despite well-described sex differences in cough severity and cough-reflex,[28–31] there are little data on differences on cough sound length. A previous study using only 234 coughs from 24 participants found shorter coughs in male.[32] Similarly, literature describing differences in cough epoch size is scarce.

Here, we show that, in this cohort, women cough-sound is significantly shorter while their epochs tend to contain more coughs. Voluntary suppression of cough by women has been proposed as a mechanism for the development of specific infections in poorly drained lung regions,[33] our findings further support this concept. Voluntary suppression and shorter coughs may contribute to suboptimal airway clearance hence driving longer cough epochs in women.

The finding of specific differences in cough length and epoch size associated with COVID-19 are also worthy of further exploration.

Among the limitations of this study, we can list that despite 400 hours of continuous audio were collected, these come from a relatively small number of patients and only 40 hours were selected for triple labelling. A proper evaluation of sex differences in cough sound length requires a much larger sample size corrected for clustering at patient level.

In summary, we describe the performance of different metrics and analysis methods to describe agreement between cough labellers. We use a robust dataset of 40 hours of continuous audio sampled from a total of over 400 hours collected from a diverse group of participants. The most intuitive way to annotate coughs is by time stamping the first explosive phase. However, this creates ambiguity as to whether sequential peaks are the same or different coughs. In contracts, we use segment-length labels and counting cough seconds and show that this metric can be used interchangeably with coughs while still capturing cough's clinically meaningful quantitative significance.

Finally, we describe sex differences in cough sound length and epoch size. These findings have implication for sexual disparities around disease progress, disease awareness and diagnosis associated with cough as a syndrome and disease transmission.

Twitter Mindaugas Galvosas @MGalvosas

Acknowledgements The authors would like to thank the participants and study team for their efforts to make this research happen.

Contributors Guarantor: CC. Conceptualisation: SGL and CC. Data curation: MR and CC. Formal analysis: JCG, MR. Funding acquisition: SGL and CC. Investigation: JCG, FC, JLDP, RM, LJ, MG and PS. Methodology: IS, MR, JCG, MG, PS, SGL and CC. Supervision: CC. Writing—original draft: IS and CC. Writing—review and editing: all authors contributed, reviewed and approved the last draft.

Funding This study was funded by the Patrick J McGovern Foundation (grant name: 'Early diagnosis of COVID-19 by using Artificial Intelligence and Acoustic Monitoring'). 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. SGL received salary support from the Fonds de Recherche en Santé Québec.

Competing interests MR, JCG, RM, LJ, MG and PS were or are employees of Hyfe and own equity in Hyfe. CCh has received consultancy fees and owns equity in Hyfe. All other authors declare no conflict of interest.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the Navarra Ethical Committee for Medical Research (CEIm), approval number PI_2021/72. All participants provided written informed consent for the study and agreed to the user policies of the Hyfe cough tracker. By design, the code linking the audio data with each participant was kept in an encrypted database accessible only to the study investigators.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to this article will be publicly available in GitHub here (https://github.com/hyfe-ai), as well as study protocol, immediately following publication, indefinitely

ORCID iDs
Mindaugas Galvosas http://orcid.org/0000-0001-7953-3309
Carlos Chaccour http://orcid.org/0000-0001-9812-050X

REFERENCES

1 Cornford CS. Why patients consult when they cough: a comparison of consulting and non-consulting patients. *Br J Gen Pract* 1998;48:1751–4.
2 Kelsall A, Decalmer S, Webster D, *et al*. How to quantify coughing: correlations with quality of life in chronic cough. *Eur Respir J* 2008;32:175–9.
3 Schmit KM, Coeytaux RR, Goode AP, *et al*. Evaluating cough assessment tools: a systematic review. *Chest* 2013;144:S0012-3692(15)48692-1:1819–26.:.
4 Hall JI, Lozano M, Estrada-Petrocelli L, *et al*. The present and future of cough counting tools. *J Thorac Dis* 2020;12:5207–23.
5 Smith J. Ambulatory methods for recording cough. *Pulm Pharmacol Ther* 2007;20:313–8.
6 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;8:00001-2022.
7 Rudd M, Song WJ, Small PM. The statistics of counting Coughs *Lung* 2022;200:531–7.
8 Kang YR, Oh JY, Lee JH, *et al*. Long-COVID severe refractory cough: discussion of a case with 6-week longitudinal cough characterization. *Asia Pac Allergy* 2022;12:e19.
9 Kulnik ST, Williams NM, Kalra L, *et al*. Cough frequency monitors: can they discriminate patient from environmental Coughs *J Thorac Dis* 2016;8:3152–9.
10 Paul IM, Wai K, Jewell SJ, *et al*. Evaluation of a new self-contained, ambulatory, objective cough monitor. *Cough* 2006;2:7.
11 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;11:e051278.
12 Clinicaltrials.gov resgitration: NCT05723159, Available: https://clinicaltrials.gov/ct2/show/NCT05723159
13 Keen EM, True EJ, Summers AR, *et al*. High-throughput Digital cough recording on a university campus: A SARS-Cov-2-negative Curated open database and operational template for acoustic screening of respiratory diseases. *Digit Health* 2022;8:20552076221097513.
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;8:00053-2022.
15 Galvosas M, Gabaldón-Figueira JC, Keen EM, *et al*. n.d. Performance evaluation of the Smartphone-based AI cough monitoring App - Hyfe cough Tracker against solicited respiratory sounds [version 1]. *F1000Res*;11:730.
16 Lee KK, Davenport PW, Smith JA, *et al*. Global physiology and pathophysiology of cough: part 1: cough Phenomenology. *CHEST Guideline and Expert Panel Report CHEST* 2021;159:282–93.
17 Miles A, Huckabee ML. Intra- and inter-Rater reliability for judgement of cough following Citric acid inhalation. *Int J Speech Lang Pathol* 2013;15:209–15.
18 Laciuga H, Brandimore AE, Troche MS, *et al*. Analysis of Clinicians' perceptual cough evaluation. *Dysphagia* 2016;31:521–30.
19 Mines D, Bacci E, Nguyen AM, *et al*. Assessment of inter- and intra-rater reliability of objective cough frequency in patients with chronic cough. ERS International Congress 2019 abstracts; September 28, 2019:suppl
20 Do W, Russell R, Wheeler C, *et al*. Performance of cough monitoring by Albus home, a Contactless and automated system for nocturnal respiratory monitoring at home. *ERJ Open Res* 2022;8:00265-2022.
21 Hirai K, Ishimaru M, Kato M, *et al*. A new method for objectively evaluating nocturnal cough in adults. *Respir Investig* 2022;60:S2212-5345(22)00002-8:400–6.:.
22 Hirai K, Tabata H, Hirayama M, *et al*. A new method for objectively evaluating childhood nocturnal cough. *Pediatr Pulmonol* 2015;50:460–8.
23 Turner RD, Bothamley GH. How to count Coughs? counting by ear, the effect of visual data and the evaluation of an automated cough monitor. *Respir Med* 2014;108:1808–15.
24 Stratil A-S, Ward C, Habte T, *et al*. Evaluating the Interrater agreement and acceptability of a new reference tool for assessing respiratory rate in children under five with cough and/or difficulty breathing. *J Trop Pediatr* 2021;67:fmab046.
25 Smith JA, Earis JE, Woodcock AA. Establishing a gold standard for manual cough counting: Video versus Digital Audio recordings. *Cough* 2006;2:6.
26 Korpás J, Sadlonová J, Vrabec M. Analysis of the cough sound: an overview. *Pulm Pharmacol* 1996;9:261–8.
27 Van Hirtum A, Berckmans D. Assessing the sound of cough towards Vocality. *Med Eng Phys* 2002;24:535–40.
28 Kastelik JA, Thompson RH, Aziz I, *et al*. Sex-related differences in cough reflex sensitivity in patients with chronic cough. *Am J Respir Crit Care Med* 2002;166:961–4.
29 Kelsall A, Decalmer S, McGuinness K, *et al*. Sex differences and predictors of objective cough frequency in chronic cough. *Thorax* 2009;64:393–8.
30 Liu W, Wu Q, Mao B, *et al*. Gender difference in the association between cough severity and quality of life among patients with Postinfectious cough. *Health Qual Life Outcomes* 2021;19:34.
31 Plevkova J, Buday T, Kavalcikova-Bogdanova N, *et al*. Sex differences in cough reflex. *Respir Physiol Neurobiol* 2017;245:S1569-9048(16)30317-2:122–9.:.
32 Olia PM, Sestini P, Vagliasindi M. Acoustic parameters of voluntary cough in healthy non-smoking subjects. *Respirology* 2000;5:271–5.
33 Reich JM, Johnson RE. Mycobacterium Avium complex pulmonary disease presenting as an isolated Lingular or middle lobe pattern. *Chest* 1992;101:1605–9.