Trial@home for children: novel non-invasive methodology for the pediatric clinical trial of the future

Kruizinga, M.D.

CHAPTER 4

Development and technical validation of a smartphone-based pediatric cough detection algorithm

Pediatr Pulmonol. 2021 Dec 29. doi: 10.1002/ppul.25801

MD Kruizinga,[1,2,3*] A Zhuparris,[1*] E Dessing,[1,2] FJ Krol,[1,3] AJ Sprij,[2] RJ Dolll,[1] FE Stuurman,[1] V Exadaktylos,[1] GJA Driessen,[2,4] AF Cohen[1,3]

**Both authors contributed equally*

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Center, Leiden, the Netherlands
4 Department of pediatrics, Maastricht University Medical Centre, Maastricht, the Netherlands

Abstract

INTRODUCTION: Coughing is a common symptom in pediatric lung disease and cough frequency has been shown to be correlated to disease activity in several conditions. Automated cough detection could provide a non-invasive digital biomarker for pediatric clinical trials or care. The aim of this study was to develop a smartphone-based algorithm that objectively and automatically counts cough sounds of children.

METHODS The training set was composed of 3,228 pediatric cough sounds and 480,780 non-cough sounds from various publicly available sources and continuous sound recordings of 7 patients admitted due to respiratory disease. A Gradient Boost Classifier was fitted on the training data, which was subsequently validated on recordings from 14 additional patients aged 0-14 admitted to the pediatric ward due to respiratory disease. The robustness of the algorithm was investigated by repeatedly classifying a recording with the smartphone-based algorithm during various conditions.

RESULTS: The final algorithm obtained an accuracy of 99.7%, sensitivity of 47.6%, specificity of 99.96%, positive predictive value of 82.2% and negative predictive value 99.8% in the validation dataset. The correlation coefficient between manual-and automated cough counts in the validation dataset was 0.97 ($p < 0.001$). The intra-and interdevice reliability of the algorithm was adequate, and the algorithm performed best at an unobstructed distance of 0.5m-1m from the audio source.

CONCLUSION: This novel smartphone-based pediatric cough detection application can be used for longitudinal follow-up in clinical care or as digital endpoint in clinical trials.

Introduction

Coughing is a physiological mechanism of the respiratory system to clear excessive secretions. It can be caused by various acute and chronic diseases, such as viral upper respiratory tract infections, bacterial infections, asthma, protracted bacterial bronchitis or tic cough, and is a common reason for parents to seek medical consultation for their children[1,2].

Several studies have shown that cough severity is correlated with disease activity in asthma and other pulmonary diseases[3-6], making cough frequency an attractive candidate biomarker for respiratory disease severity. Although coughing is traditionally quantified via self- or parent-report in the form of questionnaires, technological advances allow for more sophisticated (semi-)automatic cough monitoring methods. Indeed, several commercial and academic entities have endeavored to develop cough detection algorithms, with varying success[7]. The most notable and reliable examples are the Leicester Cough Monitor and the VitaloJak[8,9], which record sounds with a dedicated body-contact device and microphone, and subsequently use semi-automated counting methods. Several completely automated cough counting algorithms have been developed, mostly for an adult population, but none have proceeded towards widespread availability[7].

A notable disadvantage of body-contact devices is that they are inconvenient in the field of pediatrics, especially in infants and toddlers. Additionally, pediatric cough sounds exhibit more variability across different ages due to the developing respiratory- and vocal system, which can make robust detection more challenging[10]. An ideal algorithm would require no manual input, be able to monitor from a distance, and be operational on low-cost consumer devices that are readily available, such as smartphones. To date, no such algorithm has been developed in the field of pediatrics. This study aimed to develop an algorithm that objectively and automatically counts cough sounds in children based on audio features collected via a smartphone application.

Materials and Methods

Ethics and logistics

This study was conducted at the Centre for Human Drug Research (CHDR, Leiden, the Netherlands) and the Haga Teaching Hospital, Juliana Children's Hospital (the Hague, the Netherlands). Institutional review board approval was obtained (registration number

T19-080), and the study was conducted in compliance with the general data protection regulation (GDPR). The algorithm was developed as part of the CHDR MORE® system, a remote monitoring clinical trial platform. Reporting was performed in accordance with EQUATOR guidelines[11].

Data collection

A comprehensive training dataset was obtained from multiple sources. First, audio was extracted from 91 publicly available videos on YouTube that contained coughing children with an estimated age between 0-16 years old. Furthermore, 334 non-coughing audio clips were gathered from YouTube, GitHub, and the British Broadcasting Corporation (BBC) sound library. The non-coughing set contained various sounds that were expected to occur in real-life settings, such as talking, breathing, footsteps, cats, sirens, dogs barking, cars honking, snoring, glass breaking, and church clocks. Additionally, 21 children aged 0-16 and admitted due to pulmonary disease were recruited on the ward of Juliana Children's Hospital. Data of the first 7 children were used to supplement the training dataset, with a maximum of the first 150 coughs per child to avoid overrepresentation of a single subject. Remaining cough sounds of the 7 children were discarded. Data from the other 14 subjects were used as validation dataset. All audio clips were manually annotated by an investigator using Audition software (Adobe, San Jose, CA, USA). No filter was applied to remove 'silent' sections of the recording to ensure that the estimated accuracy reflects real-life conditions. As a result, the proportion of cough sounds in the validation dataset was 0.7%. The composition of the final training-and validation dataset are displayed in *Table 1*.

Audio feature extraction and selection

Audio feature were extracted from all audio clips using the OPENSMILE software (version 2.3.0, audEERING, Gilching, Germany)[12]. The software converted all audio clips into 1582 features per epoch. Epoch length was fixed at 0.5 seconds since the average cough duration in the training dataset was 0.3 seconds. The extracted features included several audio domains, such as Mel-frequency cepstral coefficients (MFCCS) and fundamental frequencies (F0) (*Supplementary Text S1*). Using manual inspection, the most robust features across multiple conditions were selected (*Supplementary Text S2*) and only these features were included in the final dataset used for algorithm development.

Table 1. Composition of training-and validation datasets

	Training dataset				Validation dataset
	YouTube (91 clips)	Various sources (334 clips)	Hospital (7 children)	Total	Hospital (14 children)
Cough sounds (n)	2,229		999	3,228	4,123
Non-cough sounds (n)	9,702	39,456	431,622	480,780	100,522
Total (n)	11,931	39,456	432,621	484,008	104,645
Cough proportion (%)*	18.5%	0%	0.2%	0.7%	0.4%
Mean cough duration (s)	0.3	-	0.3	0.3	0.3

* Proportion of 0.5 second epochs that contain cough sounds.

Algorithm development and validation

Two discriminative decision-tree based classifiers were considered for the model: Random Forest and Gradient Boost Classifier. Five-fold cross-validation was used to select the optimal features and hyperparameters for the model. The optimal classifier was selected based on the highest overall Matthew's correlation coefficient (MCC). The selected model was then used to classify all 0.5-second epochs in the validation dataset. The sensitivity, specificity, MCC, positive predictive value (PPV) and negative predictive value (NPV) were calculated for the complete validation dataset and per subject.

Initial robustness tests

Limited robustness tests were conducted to ensure the algorithm performs comparably across a range of different conditions when applied as a smartphone application. First, a 27-minute-long audio-clip was generated which included coughing- and household sounds, as well as sections with silence. The clip was subsequently played repeatedly from a speaker, while a G6 smartphone (Motorola, Chicago, IL, USA) with the CHDR MORE® application was placed in proximity. The application has incorporated OPENSMILE software and is able to calculate and transmit the generated audio features. The following conditions were tested: first, the intra-device variability was tested by repeating the assessment 7 times with the same device; second, the inter-device variability was tested by repeating the assessment 4 times with different devices of the same type;

third, the effect of device distance (0.5m, 1m and 4m) from the audio source was assessed and finally, accuracy was assessed when a small (plant and book) or large (loft bed) barrier was placed in front of the audio source and when television sounds were played in the background. Because the 0.5-second epochs from the original file and the output of the MOORE® application could not be paired, cumulative cough count plots were generated and compared across conditions.

Results

Algorithm training

The training set consisted of 3,424 0.5-second cough epochs of various sources, as well as 431,622 0.5-second non-cough epochs. The final algorithm, fitted through a Gradient Boost Classifier, achieved an accuracy of 99.6%, MCC of 73.7%, sensitivity of 99.6% and specificity of 99.9% in the training set (*Table 2*). The most important audio features the algorithm relied on were derived from the mel frequency and loudness categories (*Supplementary Figure S3*).

Table 2. Performance of the final algorithm

Parameter	Training dataset	Validation dataset
	Mean (SD) performance*	Overall performance
Accuracy	99.61% (+/- 0.13%)	99.74%
MCC	73.67% (+/- 0.16%)	62.40%
Sensitivity	99.62% (+/- 0.13%)	47.56%
Specificity	99.89% (+/- 0.09%)	99.96%
PPV	99.65% (+/- 0.08%)	82.16%
NPV	99.82% (+/- 0.02%)	99.78%

Abbreviations: MCC: Matthew's Correlation Coefficient, PPV: positive predictive value, NPV: negative predictive value. * Mean (SD) performance of 5-fold cross-validation

Algorithm validation

For validation, fourteen patients with respiratory disease aged 0-14 were recorded during a hospital admission. The median recording duration was 632 (IQR 477-775) minutes. In total, 4,123 0.5-second epochs contained coughing. The median cough count per subject was 150 (IQR 38-446). *Table 2* displays the overall accuracy of the algorithm in the validation dataset. Overall sensitivity was 47.6% and specificity was 99.96%. Due to the relatively low frequency of cough counts in the dataset, the NPV and PPV in these real-world settings was 99.78% and 82.2%, respectively. The performance of the algorithm differed between subjects. Individual patient characteristics and classification accuracies are displayed in *Table 3*. The correlation coefficient between manual cough count and automated cough count was 0.97 ($p < 0.001$, *Figure 1*).

Figure 1. Correlation manual-and automatic cough count in validation dataset. Pearson correlation between manually counted coughs and automatically detected coughs. Each dot represents an individual subject in the validation dataset.

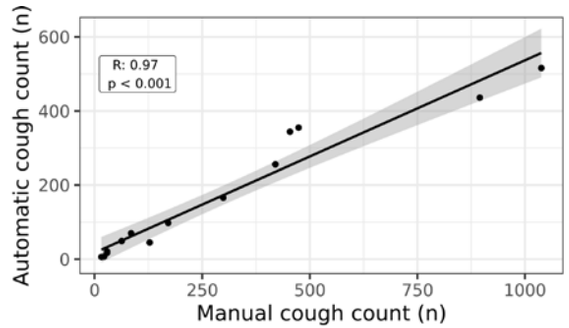


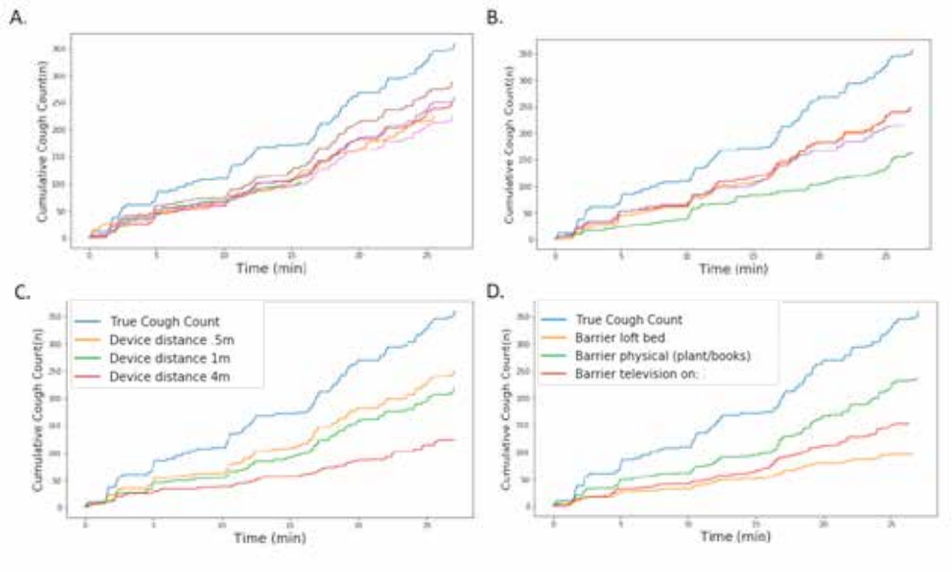
Table 3. Performance of the final algorithm in Individual subjects

Subject (#)	Age	Diagnosis	Recording duration (min)	Manual Count (n)	Algorithm count (n)	Sens.	Spec.	MCC
1	14y	Pneumonia	4	22	7	32%	100%	55%
2	4y	Wheezing	717	63	49	73%	100%	73%
3	5y	Pneumonia	237	29	21	72%	100%	85%
4	1.5y	Pneumonia	609	16	6	19%	100%	31%
5	6w	Bronchiolitis	727	85	70	58%	100%	63%
6	3y	Pneumonia	792	454	344	69%	100%	79%
7	9w	Bronchiolitis	967	895	436	34%	100%	69%
8	4y	Pneumonia/ Wheezing	497	29	17	52%	100%	88%
9	11y	Asthma	598	171	98	56%	100%	73%
10	5w	Bronchiolitis	873	1038	516	37%	100%	53%
11	2y	Pneumonia	434	474	355	70%	100%	81%
12	3y	Pneumonia	470	420	256	54%	100%	68%
13	13w	Bronchiolitis	654	128	45	34%	100%	57%
14	4y	Pneumonia	791	299	166	40%	100%	53%

Limited algorithm robustness tests

Repeated (n=7) tests with the same device and show comparable performance during each iteration (*Figure 2A*), while the inter-device variability tests show some variability in cumulative cough count across devices (*Figure 2B*). The effect of the distance of the device to the audio source was assessed (*Figure 2C*) and demonstrated comparable accuracy for 0.5 and 1m distance. The accuracy was lower when the distance of the monitoring device from the audio source was increased. Finally, the effect of a small- and large barrier was investigated, as well as the effect of ambient television sounds playing in the background (*Figure 2D*). During this test, it appeared that a small physical barrier did not impact algorithm performance, but a large physical barrier and background television sounds led to a lower cumulative cough count.

Figure 2. Performance of the algorithm under varying circumstances. A. Intra-device repeatability. Each individual line represents a different session with the same device. B. Inter-device repeatability. Each individual line represents a different session with a different device of the same type. C. Influence of device distance from the audio source. D. Influence of physical barrier or ambient background noise. In each of the panels, the light-blue line is the reference from the audio file.



Discussion

The current manuscript described the development and initial validation of a novel cough detection algorithm in pediatrics. Publicly available audio recordings were combined with real-life recordings to fit an algorithm that had excellent classification capability in the training dataset. In the validation dataset, a sensitivity of 47.6% and specificity of 99.96% was obtained, which resulted in a PPV of 82.2% and an NPV of 99.8% in these real-world conditions. There was a strong correlation between manual cough count and automatic cough count. The accuracy of the algorithm in the validation set was confirmed by several robustness tests, which repeatedly showed a cumulative cough count that was roughly half of the true cough count across various conditions. The algorithm performed best when there was a relatively unobstructed maximum distance of 0.5m-1m from the audio source.

The current sensitivity is suboptimal but does not disqualify the algorithm, and we envision the current algorithm is already suitable for application in several settings. Algorithm-derived cough count could be incorporated as (secondary) digital endpoint in pediatric pulmonary disease trials. For this application, clinical validation of cough count as digital endpoints should be performed first, focusing on demonstrating a difference between patients and healthy children, correlation of the novel endpoint with traditional endpoints or patient reported outcomes, and sensitivity to change in disease activity[15]. In addition to clinical trials, applying this algorithm in clinical care is likely to be much more reliable than patient- or parent recall regarding cough frequency[13,14]. The strong correlation between manually- and automatically-counted coughs means the algorithm can discriminate children that cough excessively from children that do not and can uncover individual trends over time, e.g. to characterize clinical recovery after a hospital admission, or to assess the effect of treatment in excessively coughing patients with persistent bacterial bronchitis. This is further supported by the very high specificity of the algorithm that is maintained in all validation tests. For example, change in nocturnal cough frequency in the case of an asthma exacerbation could be identified reliably with the current algorithm, and subsequent treatment leading to a significant decrease in nocturnal coughing will also be detectable even with the current sensitivity. In the future, algorithm output could be combined with other non-invasive assessments known to be related to pulmonary disease-activity, such as physical activity-, heart rate- and pulmonary function monitoring, as well as electronic patient reported outcome measures. Together, this could provide a holistic overview of multiple aspects of pulmonary disease-severity and quality of life[16].

Multiple research groups have developed cough detection algorithms in recent years. However, only one was developed specifically for a pediatric population[17]. Although this algorithm was not applied in a mobile device. Still, pediatric cough detection is theoretically more challenging due to changing vocal cord acoustics during various stages of development. In adults, the most widely reported cough detection devices are the Leicester cough monitor and the VitaloJak[7]. These methods have been validated in independent datasets and appear both sensitive (91-99%) and specific (99%), but the use of dedicated microphones is less user-friendly in general and the use contact-devices precludes their use in several age categories in pediatrics. Furthermore, the semi-automated counting method used by both devices remains laborious and requires training, which means that widespread use in large-scale clinical trials or in general care is not feasible. Other algorithms that count coughs automatically have reported sensitivities of 78-99% and specificities of 92-99%[7,17-22], but only a few have been applied on a smartphone[20,21]. The one that most resembles the current study is a smartphone-based algorithm developed by Barata et al, who use a convolutional neural network to classify nocturnal sounds in adult asthmatics and obtained a sensitivity of 99.9% with a specificity of 91.5%[20]. In addition, other projects are often based on data obtained in tightly controlled environments and lack validation in independent or clinical datasets[17,21,22], and may show a similar drop in accuracy during validation as was observed for the algorithm developed here. For example, the PulmoTrack® device, designed for automatic clinic-based monitoring, showed a reduced sensitivity of 26% compared to human annotation during validation in a new cohort[23].

A major advantage of the algorithm developed in this study is the conversion of raw audio into audio features on the smartphone before transmission to the study center, which ensures the privacy of participants. The automated classification is another advantage, allowing devices to analyze-and transmit cough counts in real-time. A limitation was the manual feature selection performed, which introduces a potentially subjective factor to the analysis. Furthermore, a laptop speaker was used during the initial robustness tests and using a higher quality speaker may have led to slightly different performance during these tests. However, we believe the device quality is sufficient for the purpose of testing repeatability and investigating the effects of differing conditions. During this study, a single smartphone type was used, and the observed performance may vary when other devices are used[24]. Another potential problem would arise when the sensitivity of the algorithm would be highly dependent on the underlying disease that is studied, although there is no evidence of this in the validation dataset, such factors need to be studied further during clinical validation for which we can supply the algorithm to other interested academic groups. The current algorithm is developed as a one-size-fits-all solution that can classify coughs of all pediatric patient groups and ages and that only used sound features as input variables. Although the current accuracy appears sufficient to include as digital biomarker in the applications mentioned above, the accuracy of future algorithms could improve significantly with the cost of added complexity. First, accuracy could improve by addition of additional covariates such as age, sex, and diagnosis, although this would require some user input before use. Second, the exponential increase in processing power of mobile devices could allow for the development of personalized models in the future, which would both be trained, validated, and deployed on the participants' own smartphones. A personalized classification model that is tuned to the cough characteristics of an individual could potentially be much more accurate, considering the intra-individual variability in cough sounds is assumed to be smaller compared to inter-individual variability. Future studies could also aim to quantify cough intensity, as this characteristic may have greater impact on quality of life than cough frequency[7].

Conclusion

This novel smartphone-based cough detection application is one of the first of its kind and able to count coughs in pediatric patients with a sensitivity of 47%, specificity of 99.96%, PPV of 82% and NPV of 99.8%. Although the observed sensitivity in the intended use must be improved in the future, the current algorithm may be reliable enough for longitudinal monitoring in the context of clinical trials-or care, which will be evaluated during a clinical validation process.

REFERENCES

1 Kantar A. Phenotypic presentation of chronic cough in children. J Thorac Dis. 2017;9(4):907–13.
2 Goldsobel AB, Chipps BE. Cough in the Pediatric Population. J Pediatr. 2010;156(3):352–358.e1.
3 Theodore AC, Tseng CH, Li N, Elashoff RM, Tashkin DP. Correlation of cough with disease activity and treatment with cyclophosphamide in scleroderma interstitial lung disease: Findings from the scleroderma lung study. Chest. 2012;142(3):614–21.
4 Sato R, Handa T, Matsumoto H, Kubo T, Hirai T. Clinical significance of self-reported cough intensity and frequency in patients with interstitial lung disease: A cross-sectional study. BMC Pulm Med. 2019;19(1):1–10.
5 Li AM, Tsang TWT, Chan DFY, Lam HS, So HK, Sung RYT, *et al.* Cough frequency in children with mild asthma correlates with sputum neutrophil count. Thorax. 2006;61(9):747–50.
6 Van Der Giessen L, Loeve M, De Jongste J, Hop W, Tiddens H. Nocturnal cough in children with stable cystic fibrosis. Pediatr Pulmonol. 2009;44(9):859–65.
7 Cho PSP, Birring SS, Fletcher H V., Turner RD. Methods of Cough Assessment. J Allergy Clin Immunol Pract. 2019;7(6):1715–23.
8 Birring SS, Fleming T, Matos S, Raj AA, Evans DH, Pavord ID. The Leicester Cough Monitor: Preliminary validation of an automated cough detection system in chronic cough. Eur Respir J. 2008;31(5):1013–8.
9 McGuinness K, Holt K, Dockry R, Smith J. P159 Validation of the VitaloJAK 24 Hour Ambulatory Cough Monitor. Thorax. 2012;67(Suppl 2):A131–A131.
10 Chang AB. Pediatric cough: children are not miniature adults. Lung. 2010 Jan;188 Suppl:S33–40.
11 Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, *et al.* Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. J Med Internet Res. 2016;18(12):1–10.
12 Eyben F, Schuller B. openSMILE:). ACM SIGMultimedia Rec. 2015;6(4):4–13.
13 Morey MJ, Cheng AC, McCallum GB, Chang AB. Accuracy of cough reporting by carers of Indigenous children. J Paediatr Child Health. 2013;49(3). doi: 10.1111/jpc.12118
14 Chang AB, Newman RG, Carlin JB, Phelan PD, Robertson CF. Subjective scoring of cough in children: Parent-completed vs child-completed diary cards vs an objective method. Eur Respir J. 1998;11(2):462–6.
15 Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, *et al.* Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation. Pharmacol Rev. 2020;72 (4)(October):899–909.
16 Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF. The Future of Clinical Trial Design: The Transition from Hard Endpoints to Value-Based Endpoints. 2019;371–97.
17 Amrulloh YA, Abeyratne UR, Swarnkar V, Triasih R, Setyati A. Automatic cough segmentation from non-contact sound recordings in pediatric wards. Biomed Signal Process Control. 2015;21:126–36.
18 Coyle M, Keenan D, Henderson L, Watkins M, Haumann B, Mayleben D, *et al.* Evaluation of an ambulatory system for the quantification of cough frequency in patients with chronic obstructive pulmonary disease. Cough. 2005;1(1):3–3.
19 Vizel E, Yigla M, Goryachev Y, Dekel E, Felis V, Levi H, *et al.* Validation of an ambulatory cough detection and counting application using voluntary cough under different conditions. Cough. 2010;6(1):1–8.
20 Barata F, Tinschert P, Rassouli F, Steurer-Stey C, Fleisch E, Puhan MA, *et al.* Automatic recognition, segmentation, and sex assignment of nocturnal asthmatic coughs and cough epochs in smartphone audio recordings: Observational field study. J Med Internet Res. 2020;22(7). doi: 10.2196/18082
21 Monge-Alvarez J, Hoyos-Barcelo C, Lesso P, Casaseca-De-La-Higuera P. Robust Detection of Audio-Cough Events Using Local Hu Moments. IEEE J Biomed Heal Informatics. 2019;23(1):184–96.
22 Pramono RXA, Imtiaz SA, Rodriguez-Villegas E. A cough-based algorithm for automatic diagnosis of pertussis. PLOS One. 2016;11(9):1–20.
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(12):1808–15.
24 Barata F, Kipfer K, Weber M, Tinschert P, Fleisch E, Kowatsch T. Towards device-agnostic mobile cough detection with convolutional neural networks. 2019 IEEE Int Conf Healthc Informatics, ICHI 2019. 2019;1–11.

SUPPLEMENTARY DATA

Sup. Text S1	Opensmile Audio features
Sup. Text S2	Opensmile Feature selection
Sup. Figure S2a	Example of distribution plots of each feature used during the feature selection process
Sup. Figure S3	Feature importance plot