Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Patient-level validation of a foundation-model pipeline for pediatric lung sounds: physician-annotated adventitious events are recognized, disease groups are not reliably predicted

Izzet Turkalp Akbasli

iakbasli@hacettepe.edu.tr

Hacettepe University

Oguzhan Serin

Hacettepe University

Research Article

Keywords: lung sounds, auscultation, pediatrics, foundation model, patient-level validation, cross-validation, shortcut learning, stacked generalization, calibration, explainable artificial intelligence

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11157962/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Lung-sound classifiers are usually evaluated on recording- or event-level splits, although each child contributes many recordings. We evaluated a foundation-model pipeline (PulmoVec) with the patient as the unit of partitioning and asked whether disease group can be predicted beyond age and sex.

Methods

We analyzed 19693 physician-annotated respiratory events from 736 children in the public SPRSound database. A frozen Health Acoustic Representations encoder with low-rank adapters was trained for screening (normal versus adventitious), sound pattern (normal, crackles, wheeze/rhonchi) and disease group (pneumonia, bronchial disease, normal/other); event probabilities were stacked with age, sex and auscultation site. The primary analysis was nested patient-grouped cross-validation in which every child was tested once, with a locked hold-out as secondary; comparators were the majority class, annotated event duration and demographics, and intervals came from bootstrapping patients.

Results

For physician-segmented events the area under the receiver operating characteristic curve (AUC) was 0.95 (95% CI 0.94 to 0.96) for screening and 0.95 (95% CI 0.94 to 0.96) for sound pattern, against 0.78 and 0.75 for event duration alone; the positive predictive value for adventitious events was 0.72 (95% CI 0.66 to 0.77). Disease-group prediction reached an AUC of 0.58 (95% CI 0.54 to 0.62); its accuracy (0.44) was below the majority-class rate (0.62) at event level and at patient level (0.46 versus 0.60), and age and sex carried 55.5% of the second-stage attribution for this outcome.

Conclusions

Recognition of physician-segmented adventitious events holds under patient-level validation and its discrimination is not explained by event duration or demographics, although in exploratory analyses its decisions favored short events and its positive predictive value would limit unaided use. Prediction of disease group, a clinical diagnosis, from lung-sound events alone was not demonstrated. Lung-sound studies should partition data by patient, as prior evidence indicates, and report non-acoustic comparators.

Background

Chest auscultation is the first instrument used to assess a breathless child, yet lung sounds are described inconsistently [1, 2] and examiners agree only moderately on crackles and poorly on rhonchi

[3]. Machine learning applied to electronic stethoscopes has been proposed as a way to make auscultation reproducible [4], and a meta-analysis of pediatric models pooled sensitivities and specificities near ninety percent [5], although reviews of the field stress its heterogeneity and the scarcity of external validation [6].

Such estimates depend on how data were partitioned; that meta-analysis rated overlap of children between training and test data as a risk of bias but did not stratify its pooled estimates by it. The SJTU Paediatric Respiratory Sound database (SPRSound), the pediatric resource used most often in that meta-analysis [5, 7], holds many recordings and annotated events per child, and part of its official test material comes from children who also appear in its training material [7, 8]. When one child's events fall on both sides of a split, a model can succeed by recognizing the child rather than the disease (identity confounding [9]), and record-wise validation overestimates clinical prediction performance [10, 11], including for lung sounds [12–14]. A further threat receives less attention: classifiers may exploit shortcuts [15], such as the annotated duration of the event for event-level classification [16] and age and sex for diagnosis [17].

PulmoVec is a multi-task stacking pipeline built on the Health Acoustic Representations (HeAR) foundation model [18]. Adventitious sounds are acoustic events that an examiner hears, whereas pneumonia and asthma are clinical diagnoses; whether the second can be read from the first is the question such a pipeline raises. Here we (i) estimate acoustic event classification by PulmoVec under patient-level validation against duration-only and demographics-only comparators, and (ii) test whether disease group can be predicted beyond demographics.

Methods

Design and data

The analysis plan, including the wording rules attached to the results, was written before the reported analyses were run and is kept, with a dated record of deviations, in the code repository; full methods are given in Additional file 1 (Section S1). This retrospective model development and internal validation study is reported according to TRIPOD+AI [19] (Additional file 1, Section S4). SPRSound contains recordings made with one electronic stethoscope at Shanghai Children's Medical Center, with event onsets, offsets and types annotated independently by three pediatric physicians (disagreements adjudicated by two senior physicians; no agreement statistic reported) and a recording-level diagnosis [7]. We used all annotated events with audio, excluding events without a usable label (stridor and "no event" annotations), recordings without a documented diagnosis and events with an implausible encoded age. Nothing was imputed. The base-model architecture, its training configuration and the disease grouping had been chosen during earlier development on SPRSound, in which the locked hold-out had been inspected; the analysis plan, dated 19 September 2026 and kept in the repository, therefore fixes the evaluation and its wording rules, not these pipeline choices. Patients and the public were not involved in this secondary analysis.

Outcomes

Three outcomes were modeled. *Screening* distinguished normal events from events with any adventitious sound; *sound pattern* distinguished normal events, crackles (including events with both wheeze and crackles) and wheeze/rhonchi; and *disease group* distinguished pneumonia, bronchial disease (bronchitis, asthma and bronchiolitis, grouped because their auscultatory signs overlap) and normal/other; the grouping was fixed before the analysis, and the recorded diagnoses in each group are listed in Additional file 1 (Section S2). The first two describe the event; the disease group belongs to the recording and is inherited by each of its events.

Model

Each event was cut at its annotated boundaries, centered in a 2 s clip whose margins were padded with low-level band-limited noise, band-pass filtered (100–1800 Hz) and encoded by the frozen HeAR Vision Transformer [18, 20] with low-rank adapters in its last six blocks [21]. One classifier per outcome was trained with focal loss [22], and their probabilities were combined with age, sex and auscultation site in a class-balanced LightGBM second stage [23, 24] trained only on out-of-fold probabilities and tuned with Optuna [25]. Full detail is in Additional file 1 (Sections S1 and S3).

Evaluation

The patient was always the unit of partitioning, and a program assertion verified before every fit and evaluation that no patient was shared between partitions. The primary analysis was nested patient-grouped cross-validation: five outer folds, so that each child was tested once, each with a validation partition and four-fold out-of-fold stacking inside the training patients [26, 27]. A locked hold-out, stratified by disease group, was secondary because it had been inspected once during development; it served model-specific analyses.

Comparators and statistics

We compared the full stack with the majority class and with second-stage models restricted to annotated event duration, to demographics, to the task's own base model, and to all acoustic probabilities. Because a child's events are correlated, all intervals are percentile intervals from 2000 bootstrap resamples of patients [28]; pre-specified paired contrasts used the same resampled patients, with Holm-adjusted p values (the smallest attainable value, 0.0025 after adjustment, is reported as $p < 0.003$). We report the area under the receiver operating characteristic curve (AUC; macro-averaged for three classes), accuracy, per-class measures and calibration [29], at event level and after confidence-weighted aggregation to recordings and patients. Two wording rules were fixed in advance: acoustic classification is called robust only if the lower confidence bound of its AUC exceeds the duration-only

AUC; disease-group prediction is reported as not demonstrated unless the full stack exceeds demographics (adjusted $p < 0.05$) and patient-level accuracy exceeds the majority-class rate.

Attribution and preprocessing controls

After the planned analyses, and not pre-specified, two cues belonging to the preprocessing rather than to the sound were tested by retraining one base model per acoustic outcome on the locked split: the first with the mel bands above our filter's cut-off replaced by the clip median, the second with the event not isolated but taken as a 2 s window of real surrounding audio, so that the padded fraction no longer encodes the annotated duration. Child-level count rules, a second stage without class weighting and a child-level decision curve were also examined post hoc. Second-stage models were explained with Shapley additive explanations (SHAP) [30]; base models with patch-occlusion maps on the spectrogram input [31] for events sampled by a fixed rule, checked against deletion curves [32], a classifier-randomization test [33] and integrated gradients [34].

Use of generative artificial intelligence

A generative artificial intelligence coding assistant (Claude Code, Anthropic) was used to help write and edit the analysis code, and generative artificial intelligence was also used for language editing. The analyses were specified by the authors, the code and its outputs were reviewed by the authors, and all analytical decisions were made by the authors; the study design and the interpretation of the results are the authors' own. The authors take full responsibility for the code, the analyses and the content of the manuscript.

Results

Cohort

Of 24808 annotated events, 304 had no usable label, 4802 lacked a documented diagnosis (excluded from all outcomes so that the three share one cohort; 1377 of them also lacked a participant identifier) and 9 had an implausible age, leaving 19693 events in 4985 recordings from 736 children (Fig. 1, Table 1). A child contributed a median of 18 events, so event-level partitioning places almost every child on both sides of a split.

Table 1 Cohort characteristics and the partitions of the locked hold-out

Variable	All	Train	Validation	Test	p (train vs test)
Patients	736	588	74	74	
Age (years), median (IQR)	4.9 (3.4–7.2)	4.8 (3.4–7.2)	5.5 (3.3–8.3)	5.0 (3.5–7.0)	0.702
Sex, n (%)					0.519
Female	362	290 (49.3)	32 (43.2)	40 (54.1)	
Male	374	298 (50.7)	42 (56.8)	34 (45.9)	
Disease group (patient-level), n (%)					1.000
Pneumonia	438	350 (59.5)	44 (59.5)	44 (59.5)	
Bronchial diseases	179	143 (24.3)	18 (24.3)	18 (24.3)	
Normal / Other	119	95 (16.2)	12 (16.2)	12 (16.2)	
Events	19693	15877	1913	1903	
Event type, n (%)					<0.001
Normal	16396	13200 (83.1)	1557 (81.4)	1639 (86.1)	
Fine Crackle	1692	1382 (8.7)	196 (10.2)	114 (6.0)	
Wheeze	1243	1001 (6.3)	131 (6.8)	111 (5.8)	
Rhonchi	208	172 (1.1)	18 (0.9)	18 (0.9)	
Coarse Crackle	112	89 (0.6)	10 (0.5)	13 (0.7)	
Wheeze+Crackle	42	33 (0.2)	1 (0.1)	8 (0.4)	
Auscultation site, n (%)					0.607
Left posterior (p1)	4981	4011 (25.3)	492 (25.7)	478 (25.1)	
Left lateral (p2)	5329	4335 (27.3)	497 (26.0)	497 (26.1)	
Right posterior (p3)	4562	3669 (23.1)	432 (22.6)	461 (24.2)	
Right lateral (p4)	4821	3862 (24.3)	492 (25.7)	467 (24.5)	

Values are counts (%) unless stated. Disease group is a property of the patient; event type and auscultation site are counted per event. P values compare the training and test partitions (Mann–

Whitney U test for age, chi-square test otherwise); the comparison of event types involves sparse cells.

Acoustic event classification

Under nested cross-validation the AUC was 0.948 for screening and 0.950 for sound pattern (Table 2, Fig. 2), with between-fold ranges of 0.908 to 0.961 and 0.917 to 0.967; the 74-child hold-out gave 0.965 and 0.963. Adventitious events were detected with sensitivity 0.80 and specificity 0.94, but, with 3297 adventitious events among 19693 (16.7%), the positive predictive value was 0.72 (95% CI 0.66 to 0.77); crackles were hardest (sensitivity 0.70; Additional file 1, Table S1). AUC varied little by age band, sex or site (0.928 to 0.959 for screening; Additional file 1, Section S5), but sensitivity did: from 0.89 in children younger than 3 years to 0.65 in those aged 6 to 11 years, and 0.89 to 0.53 from the shortest to the longest tertile of event duration. Discrimination remained at or above 0.90 within tertiles, but the operating point shifted: specificity rose from 0.87 to 0.98 as sensitivity fell, 71% of adventitious events lay in the shortest tertile, and a threshold matched within each tertile to the overall specificity recovered only part of the loss for long events (0.68), while long crackles were detected at 0.27 (Additional file 1, Section S2). The class-balanced second stage overstated the probability of an adventitious sound (calibration slope 0.71, intercept −1.00).

At the child level the pre-specified confidence-weighted average missed most children with an adventitious sound (sensitivity 0.28, specificity 0.99), because a child's events are mostly normal while the derived label is "any adventitious event"; calling a child positive when at least two events were predicted adventitious, a post hoc rule, gave a sensitivity of 0.79 and a specificity of 0.78 (Additional file 1, Table S3); at a prevalence of 10%, the positive predictive value would be 0.29 for children under this rule and 0.58 for single events, and in a post hoc decision-curve analysis the count rule gave a higher net benefit than treating all children from a threshold probability of about 0.11 upward and exceeded the pre-specified rule at every threshold from 0.05 to 0.40 (Additional file 1, Section S2).

Table 2 Performance of the full stacked model under patient-level validation

Outcome and level	Units	Majority-class accuracy	Accuracy	MCC	AUC, nested CV	AUC, hold-out
Screening, event	19693	0.83	0.91 (0.91–0.92)	0.71 (0.66–0.75)	0.95 (0.94–0.96)	0.97 (0.94–0.98)
Screening, patient[a]	736	0.67	0.76 (0.73–0.79)	0.44 (0.38–0.50)	0.87 (0.84–0.90)	0.89 (0.81–0.96)
Sound pattern, event	19693	0.83	0.90 (0.89–0.91)	0.69 (0.65–0.73)	0.95 (0.94–0.96)	0.96 (0.94–0.97)
Sound pattern, patient[a]	736	0.67	0.74 (0.71–0.77)	0.40 (0.34–0.46)	0.88 (0.86–0.90)	0.84 (0.76–0.91)
Disease group, event	19693	0.62	0.44 (0.40–0.48)	0.10 (0.05–0.16)	0.58 (0.54–0.62)	0.70 (0.57–0.80)
Disease group, patient	736	0.60	0.46 (0.42–0.49)	0.14 (0.09–0.19)	0.60 (0.57–0.64)	0.62 (0.50–0.74)

Nested CV (cross-validation): pooled out-of-fold predictions of the five outer folds. Intervals are percentile bootstrap intervals obtained by resampling patients. AUC is macro-averaged one-vs-rest for three-class outcomes. MCC: Matthews correlation coefficient. Macro-F1 and further metrics are given in Additional file 1, Section S6. [a] Patient-level labels of the two acoustic outcomes are derived (any adventitious event; most frequent adventitious pattern), not annotated.

Comparators and shortcuts

Annotated event duration alone reached an AUC of 0.779 for screening and 0.748 for sound pattern (Table 3, Fig. 3); the acoustic models exceeded it by 0.172 (95% CI 0.144 to 0.206) and 0.200 (95% CI 0.173 to 0.231) (adjusted $p < 0.003$), meeting the pre-specified condition for robustness, and kept AUCs of 0.902 or more within duration tertiles. Demographics alone were weakly informative (0.653 and 0.619). Stacking improved accuracy for both outcomes (by 0.018 and 0.055 against the task's own base model) and AUC for sound pattern (by 0.014), but lowered screening AUC (by 0.012); nine second-stage learners were near-identical (Additional file 1, Section S7).

Table 3 Comparator ladder: AUC at event level and paired contrasts (nested cross-validation)

Model	Screening	Sound pattern	Disease group
Event duration only	0.779 (0.740–0.813)	0.748 (0.715–0.778)	0.568 (0.543–0.593)
Demographics only	0.653 (0.578–0.715)	0.619 (0.560–0.669)	0.517 (0.474–0.563)
Own-task base model	0.960 (0.951–0.968)	0.936 (0.925–0.947)	0.593 (0.565–0.621)
All acoustic probabilities	0.952 (0.941–0.961)	0.949 (0.940–0.957)	0.594 (0.566–0.623)
Full stack	0.948 (0.936–0.958)	0.950 (0.940–0.958)	0.583 (0.543–0.624)
Δ acoustic − duration	0.172; $p < 0.003$	0.200; $p < 0.003$	0.026; $p = 0.255$
Δ full − demographics	0.295; $p < 0.003$	0.331; $p < 0.003$	0.067; $p < 0.003$
Δ full − own-task base	−0.012; $p < 0.003$	0.014; $p < 0.003$	−0.009; $p = 1.000$

All rows use the same second-stage learner and tuning protocol. Contrasts are paired patient-bootstrap differences with Holm-adjusted two-sided *p* values.

Disease-group prediction

The full stack reached an AUC of 0.583 (Table 3; outer folds 0.477 to 0.677, Additional file 1, Table S13), and its accuracy of 0.44 was below the 0.62 obtained by always predicting pneumonia; at patient level accuracy was 0.46 against 0.60 (Table 2). The signal was not zero: the full stack exceeded demographics, the contrast used by the decision rule, by 0.067 (Table 3), and the acoustic probabilities alone exceeded them by 0.078 (95% CI 0.027 to 0.127) (adjusted $p = 0.020$), and patient-level balanced accuracy was 0.41 (95% CI 0.37 to 0.45) against a chance value of one third. It was, however, far too small to be useful and did not exceed annotated event duration (adjusted $p = 0.255$). The first pre-specified condition was met and the second was not, so disease-group prediction is reported as not demonstrated. Because the class-balanced second stage trades accuracy for minority-class recall, the accuracy rule is demanding, but balanced accuracy (0.39 at event level) and the Matthews correlation coefficient (0.10 and 0.14 at event and patient level) were also low; without class weighting, in a post hoc sensitivity analysis, accuracy rose to 0.60 at event level and 0.58 at patient level, still below the majority-class rates, and no pre-specified wording rule changed. Across 42 post hoc configurations at patient and recording level (36 models fitted without class weighting and 6 unfitted mean predicted probabilities), the patient-level AUC of the fitted models reached at most 0.65 for the three groups, with accuracy within 0.01 of the majority-class rate, 0.65 for pneumonia versus other diagnoses, with accuracy of up to 0.65 against 0.60, and 0.63 for pneumonia versus healthy controls (0.645 for the unfitted mean predicted probability of pneumonia), with accuracy at the majority-class rate (recording-level AUC up to 0.66, 0.69 and 0.70, respectively; Additional file 1, Section S2). Age and sex carried 55.5% of the second-stage attribution, against 2.8% for screening (Additional file 1, Section S8).

Attribution

Removing the most attributed patches lowered the probability of the true class faster than random removal (difference in mean area under the deletion curve 0.114 for screening and 0.182 for sound pattern), but individual screening maps agreed no better with integrated gradients than with maps from a randomized classification head, so they should be read with caution (Additional file 1, Section S9). For adventitious events, 40.7% of attribution lay in one mel band of about 280 to 620 Hz (Fig. 4); for normal events in the screening model, 67.9% lay above our filter's cut-off, where only attenuated residual energy remains.

Removing either preprocessing cue did not measurably change discrimination. Against hold-out base-model AUCs of 0.967 (screening) and 0.959 (sound pattern), masking the bands above the cut-off changed the AUC by 0.002 (95% CI −0.001 to 0.007) and −0.001 (95% CI −0.008 to 0.010), and replacing the isolated clip with real surrounding audio by −0.006 (95% CI −0.013 to 0.003) and −0.006 (95% CI −0.016 to 0.014) (all $p > 0.1$); context clips admit neighboring events, so the second control may introduce another cue.

Discussion

With the patient as the unit of partitioning, recognition of physician-segmented adventitious events held: an AUC of about 0.95, far above what annotated event duration achieves, independent of demographics, and barely changed on the hold-out when either preprocessing cue was removed. Sensitivity, however, varied with age and fell for long events: discrimination stayed at or above 0.90 within duration tertiles, but the operating point shifted toward short events. Disease group was not predictable to a clinically useful degree: the pooled interval excluded chance, though one outer fold fell below 0.5, and accuracy stayed below the majority-class rate.

That record-wise evaluation inflates performance is known [10, 11], including for lung sounds [12–14]. Our outcomes differ in their exposure to that mechanism: a diagnosis is constant within a child, so a model evaluated on a record-wise split can succeed by recognizing the child [9], whereas event type varies within a child and must be heard. Disease-level labels are therefore the most exposed to this bias, and the only other application of this foundation model to SPRSound that we found does not state its unit of partitioning [35]. On SPRSound itself, the best challenge models scored 0.10 to 0.15 lower on a later blind test set, a drop that may reflect new annotators and recording conditions as well as new children (Additional file 1, Section S10).

Two explanations, which this design cannot separate, fit the disease-group result. The first is clinical. Crackles and wheeze are properties of the sound itself, and a model trained on expert annotations learned to recognize them; pneumonia is a clinical diagnosis that integrates history, fever, respiratory rate, oxygen saturation and often imaging, no single examination finding reliably establishes it in children, and against radiographic pneumonia auscultatory findings weigh less than hypoxemia and

increased work of breathing [36, 37]. In hospitalized children in low- and middle-income countries, digitally recorded crackles were not significantly associated with radiographic pneumonia [38]. The second is task construction: each event inherits its recording's diagnosis, so most acoustically normal events carry a disease label; the residual class is heterogeneous (71 of its 119 children were healthy controls); 27 children carried conflicting diagnoses between the source summary files; and the class-balanced second stage trades accuracy for minority-class recall, although removing the weighting did not change the result. DeepBreath likewise found pneumonia much harder than distinguishing healthy from pathological breathing [39], and a COVID-19 audio model was overtaken by a symptom questionnaire [17]. The post hoc child-level models, which assign each child a single label, also fell short of clinical usefulness, including for pneumonia versus healthy controls; this is consistent with the first explanation but does not establish it, because their inputs summarize three classifiers' outputs rather than the full acoustic representation. The finding concerns this labeling and these inputs, not lung sounds in general. Nor does it close the question: the acoustic probabilities carried a small signal beyond age and sex, and whether a child's aggregated lung sounds add to the clinical findings physicians already use deserves a prospective multicenter study that records oxygen saturation, respiratory rate, fever and imaging alongside the recordings.

These estimates must not be read as performance on unsegmented recordings: the models classify events whose boundaries physicians drew, a task no clinician performs, and a multi-country pediatric study without such boundaries reported balanced accuracies near seventy percent [40]. The models could serve as a second reader or teaching aid on annotated recordings, or as the classification stage behind an automatic event detector; it is that combined, child-level system that needs prospective evaluation. At the child level the pre-specified averaging rule missed most affected children; counting events did far better, but that rule was chosen post hoc and needs confirmation. Stacking, the pipeline's defining design feature, added little: it improved accuracy but slightly lowered screening AUC, so discrimination rests on the adapted foundation model.

Limitations include a single center and device, no external validation, and dependence on expert segmentation. Nor was a cross-release validation performed; it is feasible, but the releases differ sharply (adventitious events made up 24.1% of annotated events in the 2022 release and 5.3% in the 2024 release, with median annotated durations of 1.2 and 2.0 s), so its result would mix new children with differences in annotation and recording (Additional file 1, Section S10). The hold-out had been inspected before the analysis plan was written, so nested cross-validation is primary; its pooled bootstrap intervals are somewhat narrow, which is why fold ranges are given. Child-level acoustic labels are derived, the rarest event types were few, and probabilities need recalibration. The single-model ablations exclude large effects, not small ones. No sample size was calculated; the 736 children bound the disease-group AUC to within about ± 0.04, enough to exclude an AUC above 0.64 for the pre-specified model; post hoc configurations reached 0.65 at patient level and 0.70 at recording level, with upper confidence bounds up to 0.71 and 0.75. The 1800 Hz cut-off may attenuate high-frequency components of crackles, the hardest class. The value of acoustic classifiers in primary care and low-resource settings is untested, and replication and deployment depend on access to the HeAR weights, which are not redistributable.

Researchers should partition by patient and assert it in code, state the unit of partitioning in the abstract, and report demographics-only and duration-only comparators.

Conclusions

A foundation-model pipeline recognizes physician-segmented adventitious lung-sound events in children with high discrimination under patient-level validation, with discrimination not explained by event duration or demographics, though its preference for short events, its positive predictive value and the child-level sensitivity of the pre-specified averaging rule, would limit its use unaided. Prediction of disease group from lung-sound events was not demonstrated, consistent with diagnoses that rest on clinical findings beyond auscultation; whether lung sounds add to those findings remains to be tested. Patient-level partitioning, which prior work shows to be necessary, and non-acoustic comparators, which here reached an AUC near 0.78, should be standard in lung-sound studies.

Abbreviations

AUC: area under the receiver operating characteristic curve; CI: confidence interval; CV: cross-validation; HeAR: Health Acoustic Representations; MCC: Matthews correlation coefficient; SHAP: Shapley additive explanations; SPRSound: SJTU Paediatric Respiratory Sound database; TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, artificial intelligence extension.

Declarations

Ethics approval and consent to participate

This study is a retrospective secondary analysis of the publicly available, fully de-identified SPRSound database, which was collected at Shanghai Children's Medical Center with approval from its Institutional Review Board (approval No. SCMCIRB-K2019056-1) [7]. Because only publicly available, de-identified data were used, no additional ethics committee approval or informed consent was sought; the study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the SPRSound repository, https://github.com/SJTU-YONGFU-RESEARCH-GRP/SPRSound [7, 8]. Participant-level partition

assignments are provided in the PulmoVec repository (results_clean/nested_cv/patient_fold_assignment.csv and results_clean/holdout_patient_assignment.csv). Project name: PulmoVec. Project home page: https://github.com/turkalpmd/PulmoVec. Archived version: commit 013a6e5 of the repository above. Operating system: Linux. Programming language: Python 3.12. Other requirements: PyTorch 2.9, LightGBM, Optuna, SHAP; the HeAR encoder must be obtained from Google under the Health AI Developer Foundations terms [41] and is not redistributed. License: MIT. Restrictions to use by non-academics: those of the HeAR terms of use.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' contributions

ITA conceived the study, designed the methodology, developed the software, performed the analysis, curated the data, prepared the figures and led the writing of the manuscript. OS contributed to validation and supervision and revised the manuscript critically. Both authors read and approved the final manuscript.

Acknowledgements

We thank the SPRSound investigators for making their data public. A preprint of this article is available on arXiv (arXiv:2603.15688).

References

1. Pasterkamp H, Brand PLP, Everard M, Garcia-Marcos L, Melbye H, Priftis KN. Towards the standardisation of lung sound nomenclature. Eur Respir J. 2016;47:724–32. https://doi.org/10.1183/13993003.01132-2015.
2. Bohadana A, Izbicki G, Kraman SS. Fundamentals of Lung Auscultation. N Engl J Med. 2014;370:744–51. https://doi.org/10.1056/nejmra1302901.
3. Florin TA, Ambroggio L, Brokamp C, Rattan MS, Crotty EJ, Kachelmeyer A, et al. Reliability of Examination Findings in Suspected Community-Acquired Pneumonia. Pediatrics.

2017;140:e20170310. https://doi.org/10.1542/peds.2017-0310.

4. Huang D-M, Huang J, Qiao K, Zhong N-S, Lu H-Z, Wang W-J. Deep learning-based lung sound analysis for intelligent stethoscope. Military Med Res. 2023;10:44. https://doi.org/10.1186/s40779-023-00479-3.
5. Park JS, Park S-Y, Moon JW, Kim K, Suh DI. Artificial Intelligence Models for Pediatric Lung Sound Analysis: Systematic Review and Meta-Analysis. J Med Internet Res. 2025;27:e66491. https://doi.org/10.2196/66491.
6. Landry V, Matschek J, Pang R, Munipalle M, Tan K, Boruff J, et al. Audio-based digital biomarkers in diagnosing and managing respiratory diseases: a systematic review and bibliometric analysis. Eur Respiratory Rev. 2025;34:240246. https://doi.org/10.1183/16000617.0246-2024.
7. Zhang Q, Zhang J, Yuan J, Huang H, Zhang Y, Zhang B, et al. SPRSound: Open-Source SJTU Paediatric Respiratory Sound Database. IEEE Trans Biomed Circuits Syst. 2022;16:867–81. https://doi.org/10.1109/tbcas.2022.3204910.
8. SJTU-YONGFU-RESEARCH-GRP, SPRSound. Open-Source SJTU Paediatric Respiratory Sound Database (GitHub repository). 2025. https://github.com/SJTU-YONGFU-RESEARCH-GRP/SPRSound. Accessed 19 Sep 2026.
9. Chaibub Neto E, Pratap A, Perumal TM, Tummalacherla M, Snyder P, Bot BM, et al. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. npj Digit Med. 2019;2:99. https://doi.org/10.1038/s41746-019-0178-x.
10. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. GigaScience. 2017;6:gix019. https://doi.org/10.1093/gigascience/gix019.
11. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4:100804. https://doi.org/10.1016/j.patter.2023.100804.
12. Pham L, Phan H, Palaniappan R, Mertins A, McLoughlin I. CNN-MoE Based Framework for Classification of Respiratory Anomalies and Lung Disease Detection. IEEE J Biomedical Health Inf. 2021;25:2938–47. https://doi.org/10.1109/jbhi.2021.3064237.
13. Kim J-W, Toikkanen M, Koo H, Kim YT, Kwon D, Kim K. Meta-Ensemble Learning with Diverse Data Splits for Improved Respiratory Sound Classification. arXiv. 2026. https://doi.org/10.48550/arXiv.2604.24096
14. Huang D, Wang L, Wang W. A Multi-Center Clinical Trial for Wireless Stethoscope-Based Diagnosis and Prognosis of Children Community-Acquired Pneumonia. IEEE Trans Biomed Eng. 2023;70:2215–26. https://doi.org/10.1109/tbme.2023.3239372.
15. Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nat Mach Intell. 2020;2:665–73. https://doi.org/10.1038/s42256-020-00257-z.
16. Rocha BM, Pessoa D, Marques A, Carvalho P, Paiva RP. Automatic Classification of Adventitious Respiratory Sounds: A (Un)Solved Problem? Sensors. 2020;21:57. https://doi.org/10.3390/s21010057

17. Coppock H, Nicholson G, Kiskin I, Koutra V, Baker K, Budd J, et al. Audio-based AI classifiers show no evidence of improved COVID-19 screening over simple symptoms checkers. Nat Mach Intell. 2024;6:229–42. https://doi.org/10.1038/s42256-023-00773-8.
18. Baur S, Nabulsi Z, Weng W-H, Garrison J, Blankemeier L, Fishman S, et al. HeAR – Health Acoust Representations arXiv. 2024. https://doi.org/10.48550/arXiv.2403.02522.
19. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.
20. Google. HeAR (Health Acoustic Representations) model card, version 1.0.0. 2023. https://huggingface.co/google/hear. Accessed 19 Sep 2026.
21. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S et al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv. 2021. https://doi.org/10.48550/arXiv.2106.09685
22. Lin T-Y, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. In: 2017 IEEE international conference on computer vision (ICCV). IEEE; 2017. pp. 2999–3007. https://doi.org/10.1109/iccv.2017.324
23. Wolpert DH. Stacked generalization. Neural Netw. 1992;5:241–59. https://doi.org/10.1016/s0893-6080(05)80023-1.
24. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in neural information processing systems. 2017.
25. Akiba T, Sano S, Yanase T, Ohta T, Koyama M, Optuna. A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. ACM; 2019. pp. 2623–31. https://doi.org/10.1145/3292500.3330701
26. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. https://doi.org/10.1186/1471-2105-7-91.
27. Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J Mach Learn Res. 2010;11:2079–107.
28. Rutter CM. Bootstrap estimation of diagnostic accuracy with patient-clustered data. Acad Radiol. 2000;7:413–9. https://doi.org/10.1016/s1076-6332(00)80381-5.
29. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.
30. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67. https://doi.org/10.1038/s42256-019-0138-9.
31. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. Computer vision – ECCV 2014. Springer International Publishing; 2014. pp. 818–33. https://doi.org/10.1007/978-3-319-10590-1_53.

32. Petsiuk V, Das A, Saenko K. RISE: Randomized Input Sampling for Explanation of Black-box Models. arXiv. 2018. https://doi.org/10.48550/arXiv.1806.07421
33. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps. In: Advances in neural information processing systems. 2018.
34. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: Proceedings of the 34th international conference on machine learning. 2017. pp. 3319–28.
35. Ehtesham A, Kumar S, Singh A, Khoei TT. Pediatric Asthma Detection with Googles HeAR Model: An AI-Driven Respiratory Sound Classifier. arXiv. 2025. https://doi.org/10.48550/arXiv.2504.20124
36. Rambaud-Althaus C, Althaus F, Genton B, D'Acremont V. Clinical features for diagnosis of pneumonia in children younger than 5 years: a systematic review and meta-analysis. Lancet Infect Dis. 2015;15:439–50. https://doi.org/10.1016/S1473-3099(15)70017-4.
37. Shah SN, Bachur RG, Simel DL, Neuman MI. Does this child have pneumonia? The rational clinical examination systematic review. JAMA. 2017;318:462–71. https://doi.org/10.1001/jama.2017.9039.
38. McCollum ED, Park DE, Watson NL, Fancourt NSS, Focht C, Baggett HC, et al. Digital auscultation in PERCH: associations with chest radiography and pneumonia mortality in children. Pediatr Pulmonol. 2020;55:3197–208. https://doi.org/10.1002/ppul.25046.
39. Heitmann J, Glangetas A, Doenz J, Dervaux J, Shama DM, Garcia DH, et al. DeepBreath—automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries. npj Digit Med. 2023;6:104. https://doi.org/10.1038/s41746-023-00838-3.
40. Kala A, Elhilali M. Multi-Stage Respiratory Sound Analysis: Confidence-Driven Wheeze and Crackle Detection. IEEE Trans Biomed Eng. 2026;73:1696–704. https://doi.org/10.1109/tbme.2025.3613489.
41. Google. Health AI Developer Foundations Terms of Use. 2024. https://developers.google.com/health-ai-developer-foundations/terms. Accessed 19 Sep 2026.

Figures

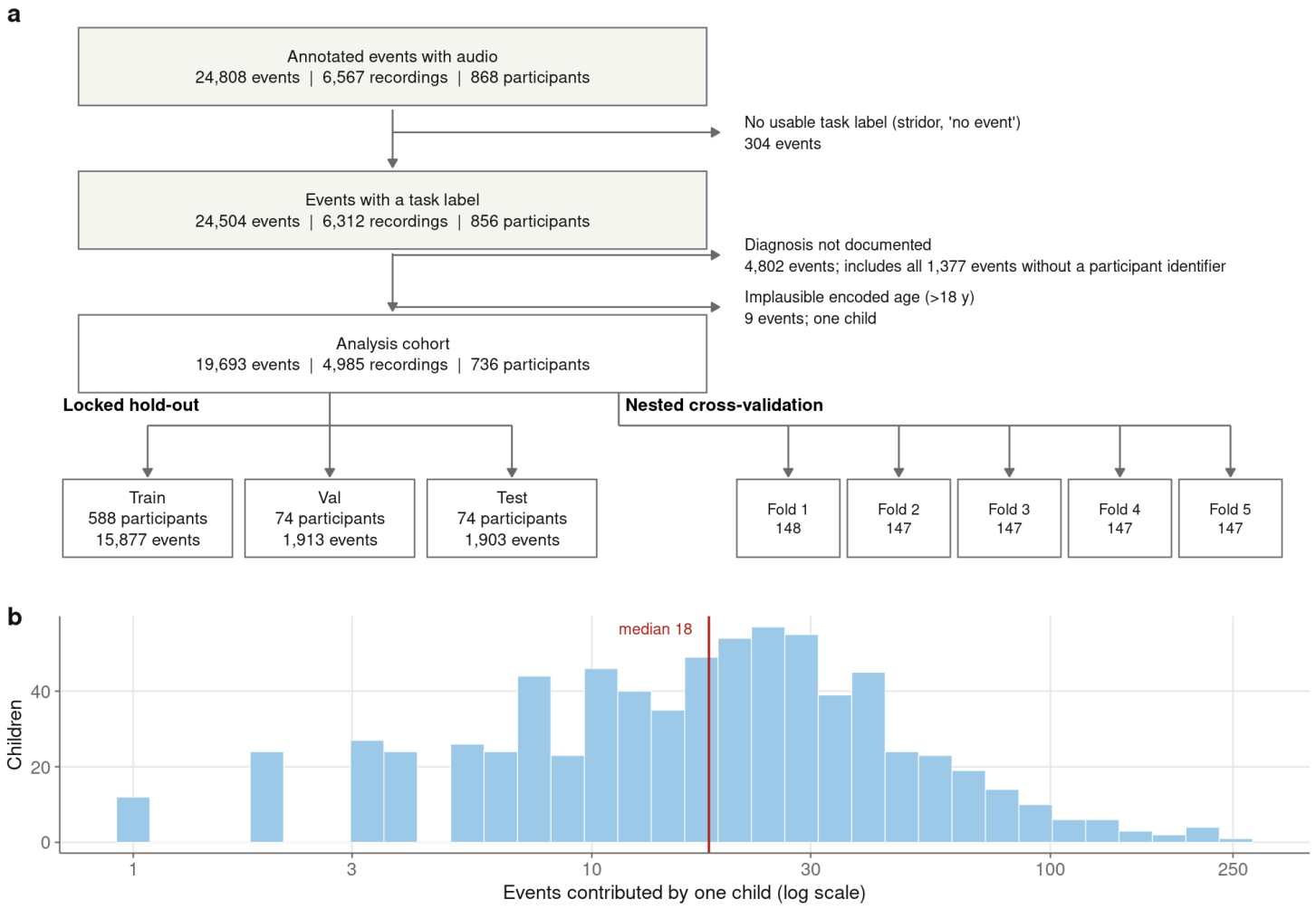


Figure 1

Flow of events, recordings and participants, and the number of events per child. a Exclusions are applied to events; counts of recordings and participants are given at each stage. The analysis cohort is then partitioned in two ways: a locked hold-out, used as the secondary analysis, and a five-fold nested patient-grouped cross-validation, used as the primary analysis, in which each participant is tested exactly once. Within an outer fold the remaining participants form a validation partition and a training partition, the latter carrying four-fold patient-grouped out-of-fold stacking. **b** Distribution of the number of annotated events contributed by each child (logarithmic scale; red line, median). Because one child contributes many events, a split drawn on events leaves almost every child on both sides of it, whereas a split drawn on children cannot.

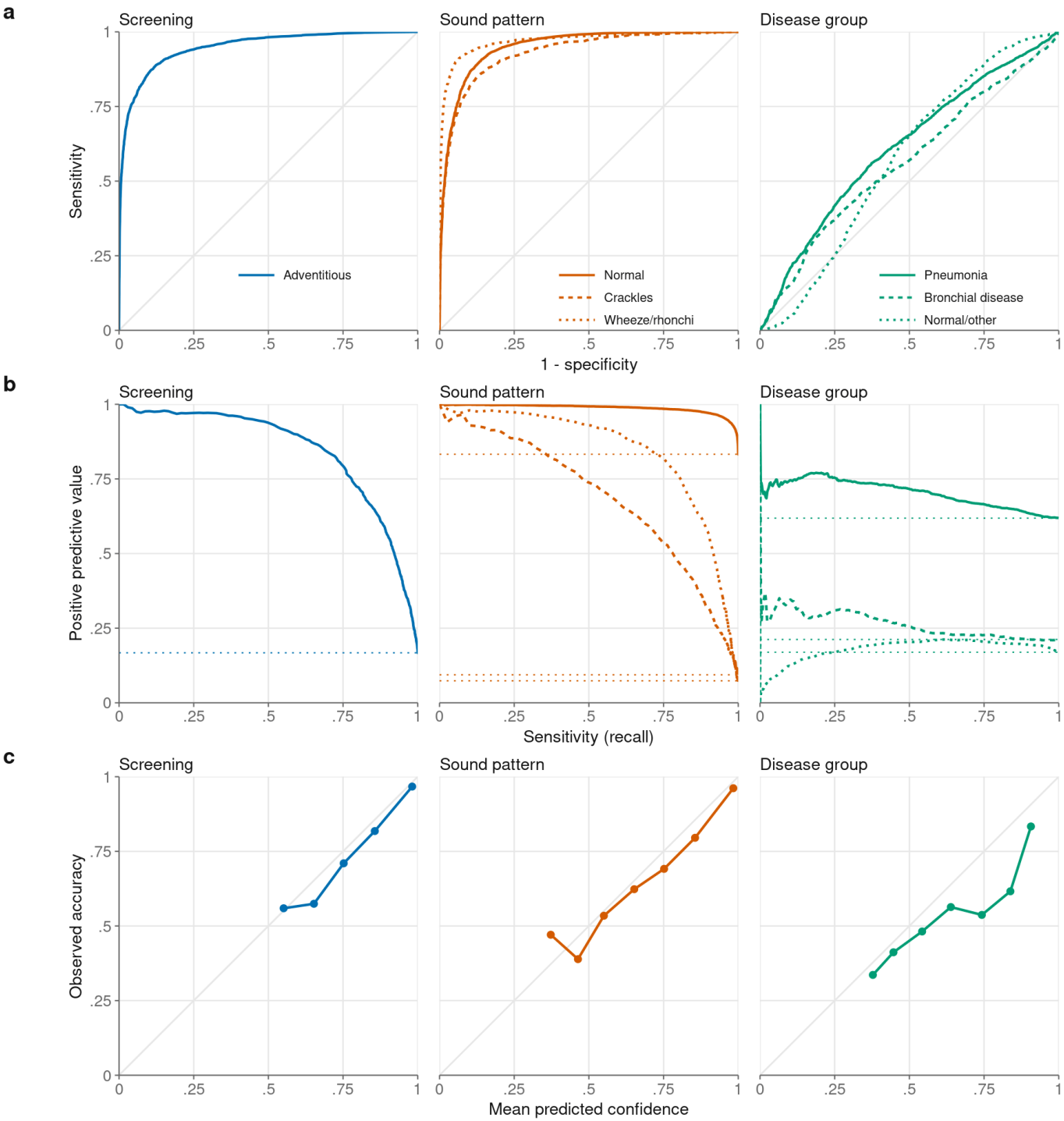


Figure 2

Discrimination and calibration under patient-level validation. Pooled out-of-fold predictions of the nested cross-validation for the full stacked model. Color identifies the outcome; line type identifies the class within an outcome, keyed in the panels of **a**. **a** Receiver operating characteristic curves; the corresponding areas under the curve with intervals are given in Table 2. **b**Precision–recall curves; the dotted horizontal lines mark class prevalence, which is the performance of an uninformative model. **c** Reliability curves: predicted confidence against observed accuracy, in ten bins. Points below the diagonal indicate over-confidence, which is what the class-balanced second stage produces for screening.

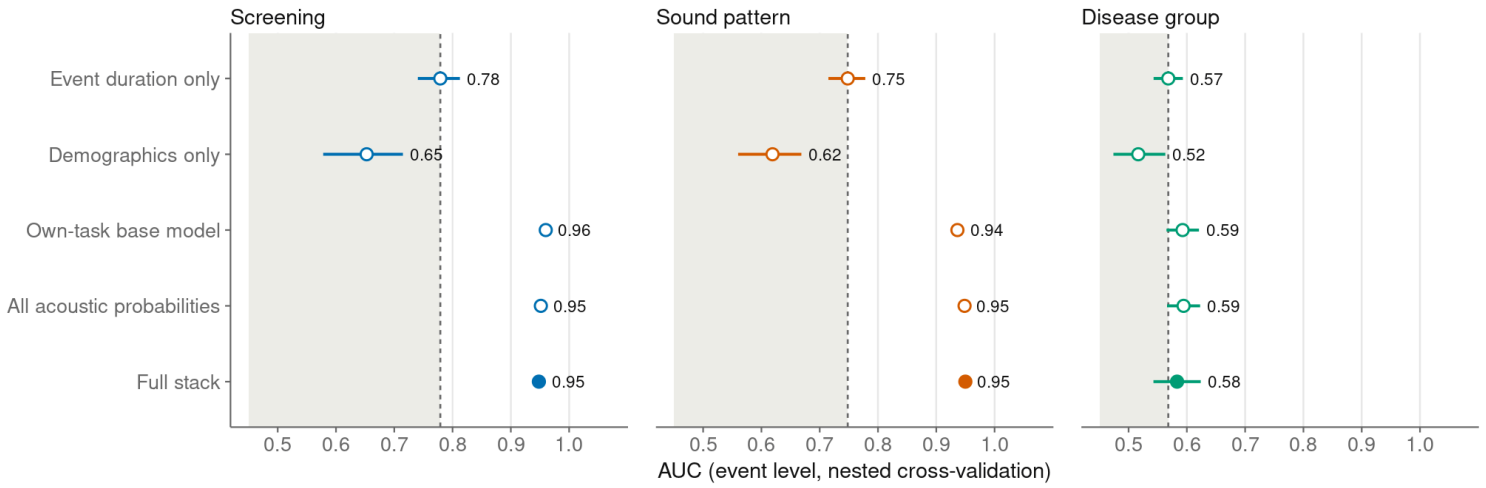


Figure 3

What the acoustic models add over two shortcuts. Event-level area under the curve of second-stage models restricted to successive feature sets, from the nested cross-validation; bars are patient-bootstrap intervals, filled circles mark the full stack. All rows use the same learner and tuning protocol. The shaded region and the dashed line mark the performance reached by the annotated event duration alone, a feature that is a property of the annotation rather than of the sound. The two acoustic outcomes lie far to the right of that line; for disease group the whole ladder lies within or at the edge of it, and the lower bound of the full stack does not clear it.

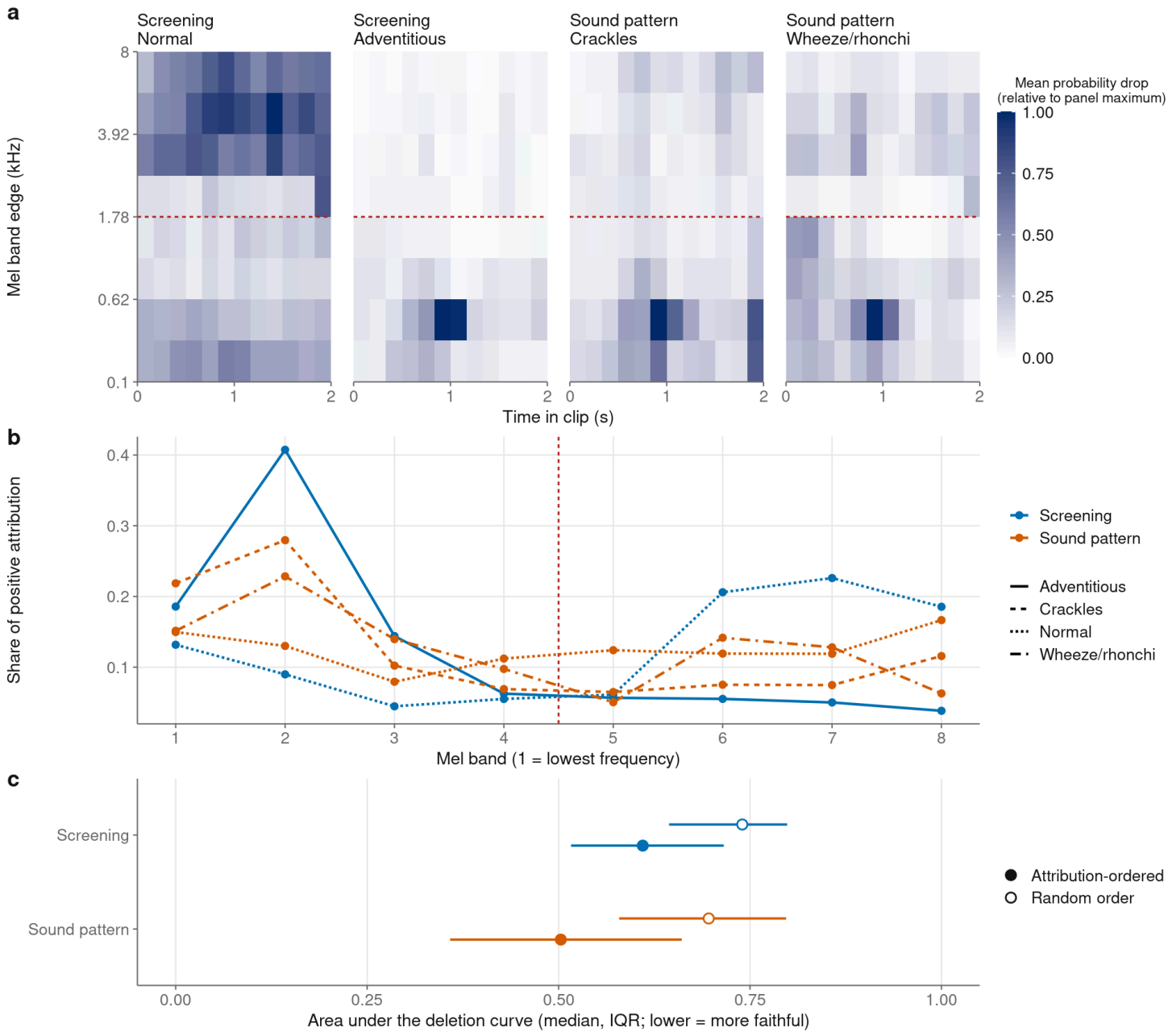


Figure 4

Model-derived attribution for the two acoustic outcomes. a Class-mean occlusion maps over the twelve-by-eight patch grid of the encoder input (time by mel band, with band edges in kilohertz) for correctly classified hold-out events, scaled to the maximum of each panel; the dashed line marks the upper cut-off of the band-pass filter applied during preprocessing, above which only attenuated residual energy remains. **b** Share of positive attribution by mel band; color identifies the outcome and line type the class. Attribution for adventitious events concentrates in a low band, whereas attribution for normal events in the screening model lies largely above the filter cut-off; retraining without that region did not change discrimination (see Results). **c** Area under the deletion curve, median and interquartile range, when patches are removed in order of attribution and in random order; a lower value for attribution-ordered removal indicates that the maps identify regions the model actually uses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx