Beyond Classification: A Cough Regression Benchmark for Respiratory Acoustic Foundation Models

Anonymous Authors[1]

Abstract

Respiratory acoustic foundation models (FMs) excel at cough classification, yet their ability to predict continuous health quantities from cough audio remains largely unexplored, despite the clinical value of passive age, BMI, and disease-probability estimation in settings where physical measurements are unavailable. We introduce the multi-model, multi-target cough regression benchmark evaluating five FMs (OPERA-CT, OPERA-CE, OPERA-GT, HEAR, M2D+RESP) across six targets on three datasets under subject-disjoint protocols, comparing linear, MLP-small, and full MLP regression heads. MLP-small beats the mean-predictor baseline on all tasks and linear probing in 23 of 30 model $\times$ task cases, with full MLP overfitting on small clinical data but recovering on larger sets, revealing a dataset-size $\times$ head-capacity trade-off. HEAR leads within-dataset age regression (CIDRZ: 10.29 yr, Coswara: 9.12 yr MAE), OPERA-GT consistently outperforms OPERA-CT on age regression across all three datasets extending a generative-pretraining advantage from breath to cough, and HEAR and M2D+RESP reach near-full performance at $N = 50$ samples while OPERA models require $N = 400$. Cross-dataset transfer is strongly asymmetric as large diverse data generalises to small clinical populations (CoughVID $\rightarrow$ CIDRZ: -0.17 yr) but not vice versa (CIDRZ $\rightarrow$ Coswara: $+2.43$ yr, $+26.6\%$).

1. Introduction

Cough acoustics encode physiological state beyond categorical disease labels, with spectrotemporal features reflecting airway geometry, respiratory muscle strength, and mucosal viscosity, all of which covary quantitatively with age, body composition, and disease severity (Sharan et al., 2018; Xu et al., 2022; Rudraraju et al., 2020). In low- and middle-income countries (LMICs) where respiratory disease burden is highest (World Health Organization, 2021), birth records, weighing equipment, and radiological capacity are often unavailable, making passive cough-based estimates of age, BMI, and X-ray abnormality probability actionable proxies for triage and dosing decisions.

Foundation models (FMs) pretrained on large unlabeled audio corpora learn task-agnostic embeddings that transfer efficiently via linear probing (Zhang et al., 2024; Baur et al., 2024; Niizumi et al., 2025), reducing the labeled-data burden in clinical audio AI. The three leading respiratory FM families, OPERA (Zhang et al., 2024), HEAR (Baur et al., 2024), and M2D+RESP (Niizumi et al., 2025), have been benchmarked extensively on classification, but their regression capability is almost entirely uncharacterised.

The HEAR paper (Baur et al., 2024) evaluates age and BMI regression from cough on the CIDRZ and Coswara (Bhattacharya et al., 2023) datasets, but restricts the analysis to a single model with a fixed linear probe and per-device splits, with no multi-model comparison, non-linear head evaluation, or cross-dataset generalisation. The OPERA benchmark (Zhang et al., 2024) includes regression tasks (spirometry estimation) exclusively from deep-breath and vowel sounds, not from cough. M2D+RESP (Niizumi et al., 2025) has never been evaluated on cough regression task.

We address these gaps with five contributions.

1. **Multi-model, multi-target cough regression benchmark** evaluating five FMs across six targets (age, BMI, X-ray abnormality, TB probability) on three datasets under subject-disjoint protocols, with MAE reported alongside the mean-predictor baseline (MAD).
2. **Regression head comparison** showing MLP-small wins 23 of 30 model $\times$ task combinations. Full MLP overfits on small clinical data ($+0.53$ yr for M2D+RESP on CIDRZ) but recovers on larger ones.
3. **Generative pretraining advantage** where OPERA-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

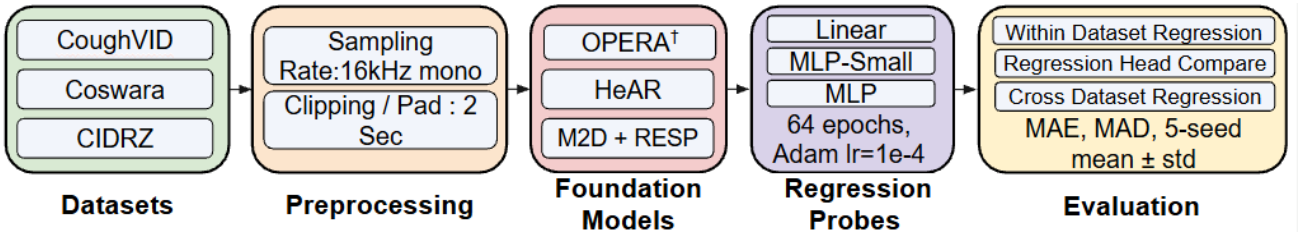

Figure 1. Cough regression benchmark pipeline. All audio is resampled to 16 kHz mono and padded/trimmed to 2 s. Five frozen encoders feed three regression probes and outputs cover three evaluation regimes. OPERA† comprises OPERA-CT, OPERA-CE, and OPERA-GT.

GT outperforms OPERA-CT on age regression across all three datasets (CIDRZ 10.49 vs. 10.52 yr, Coswara 10.16 vs. 10.25 yr, CoughVID 9.62 vs. 9.79 yr), extending the breath-regression finding of Zhang et al. (2024) to cough.

4. **Cross-dataset transfer asymmetry** where large web-collected data generalises to small clinical populations (CoughVID → CIDRZ -0.17 yr) while the reverse fails (CIDRZ → Coswara $+2.43$ yr, $+26.6\%$).

5. **Low-data regime analysis** showing HEAR and M2D+RESP reach near-full performance at $N = 50$ samples while OPERA models require $N = 400$, indicating pretraining corpus diversity determines low-data regression performance.

The rest of this paper is organized as follows. Section 2 describes the benchmark design. Section 3 presents results and Section 4 summarises findings and future directions.

2. Benchmark Design

We evaluate five frozen FMs on three cough datasets under a unified pipeline (Figure 1), resampling all audio to 16 kHz mono and padding or trimming to 2 s, with embeddings extracted once and shared across all regimes.

2.1. Datasets and Tasks

Table 1 summarises the three datasets and six regression targets. Subject-disjoint splits (64/16/20%) are used for CIDRZ and Coswara; CoughVID reuses the official UUID-level split (train = 3050, val = 1019, test = 2789).

CIDRZ (Baur et al., 2024) ($N = 1049$) contains smartphone-recorded volitional coughs from patients at a TB clinic in Zambia, with labels from clinical assessments covering age (37.1 ± 12.9 yr), BMI (21.6 ± 5.3 kg/m^2), chest X-ray abnormality probability (0.46 ± 0.35), and TB probability (0.62 ± 0.24).HEAR's pretraining corpus may include CIDRZ recordings, potentially inflating within-dataset CIDRZ performance; cross-dataset results are less affected.

Coswara (Bhattacharya et al., 2023) ($N = 2560$) consists of shallow-cough recordings collected remotely via a web application across India during the COVID-19 pandemic, with self-reported age labels (35.1 ± 13.9 yr, range 7–87).

CoughVID (Orlandic et al., 2021) ($N = 6858$) comprises volitional cough recordings submitted globally via smartphone with continuous age labels (34.5 ± 12.7 yr, range 5–97) and the largest training set (3,050 samples).

Table 1. Dataset and task summary. Prob. = prob. in $[0, 1]$.

Dataset	Task	Unit	N
CIDRZ (Baur et al., 2024) Zambia	Age	yr	1049
	BMI	kg/m^2	
	X-ray Abn.	prob.	
	TB probability	prob.	
Coswara (Bhattacharya et al., 2023) India	Age	yr	2560
CoughVID (Orlandic et al., 2021) Global	Age	yr	6858

Table 2. Foundation models. All receive 2 s at 16 kHz. RC = respiratory clips. † 8.18 s positional grid; padded to 2 s.

Model	Architecture	Dim	Pretraining
OPERA-CT	Contrastive Trans.	768	136K RC
OPERA-CE	Contrastive CNN	1280	136K RC
OPERA-GT	Generative MAE†	384	136K RC
HEAR	ViT-L MAE	512	313M health
M2D+RESP	Masked Mod. + Resp	3840	AudioSet+RC

2.2. Foundation Models

We evaluate five frozen FMs spanning three pretraining paradigms: respiratory-specific SSL (OPERA), large-scale health audio masked autoencoding (HEAR), and general audio masked modelling with respiratory fine-tuning (M2D+RESP). Table 2 summarises the five encoders.

OPERA-CT and OPERA-CE (Zhang et al., 2024) are contrastive models on 136K respiratory clips, differing in architecture (Transformer vs. EfficientNet CNN) and dimension (768-d vs. 1280-d). OPERA-GT (Zhang et al., 2024) is a

Table 3. Within-dataset MAE (MLP-small, 5-seed mean ± std). MAD = mean absolute deviation (naive mean-predictor baseline); best/MAD = ratio of best MAE to MAD (lower is better; <0.90 indicates meaningful signal above chance). **Bold** = best MAE per row.

Task	Unit	MAD	OPERA-CT	OPERA-CE	OPERA-GT	HEAR	M2D+RESP	best/MAD
CIDRZ Age	yr	10.35	10.52 ±0.08	10.51 ±0.09	10.49 ±0.07	**10.29** ±0.04	10.40 ±0.05	0.99
CIDRZ BMI	kg/m²	3.74	**3.60** ±0.01	**3.60** ±0.01	3.67 ±0.01	**3.60** ±0.02	3.63 ±0.02	0.96
CIDRZ Abn	prob	0.325	0.327 ±0.001	0.325 ±0.001	**0.316** ±0.001	0.328 ±0.001	0.320 ±0.004	0.97
CIDRZ TB	prob	0.205	0.189 ±0.001	0.191 ±0.000	0.190 ±0.000	**0.188** ±0.001	0.192 ±0.001	0.92
Coswara Age	yr	11.31	10.24 ±0.02	10.44 ±0.01	10.16 ±0.04	**9.12** ±0.07	9.58 ±0.06	**0.81**
CoughVID Age	yr	10.29	9.79 ±0.01	9.88 ±0.02	9.62 ±0.03	**9.61** ±0.02	9.79 ±0.02	0.93

generative masked autoencoder on the same corpus with an 8.18 s positional grid and zero-padded inputs. HEAR (Baur et al., 2024) is a ViT-L masked autoencoder pretrained on 313M health audio clips, the largest corpus and the only one covering clinical audio beyond respiratory sounds. M2D+RESP (Niizumi et al., 2025) combines masked spectrogram prediction on AudioSet with respiratory fine-tuning, with features mean-pooled to a single 3840-d vector per clip. All encoders are kept frozen and embeddings are extracted once, reused across all heads and evaluation regimes.

2.3. Regression Heads and Training

We compare three heads applied to frozen embeddings. **Linear** ($d_{\text{feat}} \to 1$) is the standard linear probing baseline (Zhang et al., 2024; Baur et al., 2024). **MLP-small** uses a 256-unit bottleneck with ReLU and 0.3 dropout ($d_{\text{feat}} \to 256 \to 1$), making it embedding-size-agnostic and comparable across all FMs. **MLP** is full-width ($d_{\text{feat}} \to d_{\text{feat}} \to 1$), yielding ∼15M hidden parameters for M2D+RESP, expected to overfit on small datasets and recover on larger ones. All heads use Adam (lr = 10^{-4}, L2 = 10^{-5}, batch 64), LR decay 0.97/epoch, MSE loss, and early stopping (patience 10) on validation MAE for up to 64 epochs. Results are mean ± std over 5 seeds.

2.4. Evaluation Regimes and Metrics

We report MAE in natural units (years, kg/m^2, probability) alongside the mean absolute deviation (MAD) of each label distribution as a naive mean-predictor baseline. MAE < MAD confirms learning above chance and best MAE/MAD quantifies signal strength, since a model collapsing to the population mean can appear effective with no patient-level prediction. Lower MAE is better throughout.

Within-dataset regression evaluates all six targets on held-out test splits (Table 3). **Regression head comparison** covers 90 model × task × head combinations to characterise the dataset-size × head-capacity interaction (Table 4). **Cross-dataset transfer** trains MLP-small on one dataset and evaluates on all others without adaptation across six age transfer conditions (Table 5)

3. Results

3.1. Within-Dataset Regression

Table 3 reports within-dataset MAE for MLP-small alongside the mean-predictor baseline (MAD, computed directly from each label distribution). All models exceed the baseline on every task, but the margin varies substantially: Coswara age shows the strongest signal (best/MAD = 0.81, HEAR −0.46 yr over M2D+RESP), while CIDRZ age remains within 1% of the baseline (best/MAD = 0.99), indicating that current FM embeddings encode limited age-specific information for this clinical cohort under linear probing. Clinical score targets show near-zero inter-model spread (TB: ≤0.004, X-ray: ≤0.012 MAE), consistent with a shared representational ceiling across all FMs.

3.2. Generative vs. Contrastive Pretraining

OPERA-GT consistently outperforms OPERA-CT on age regression across all three datasets: CIDRZ 10.49 vs. 10.52 yr, Coswara 10.16 vs. 10.25 yr, and CoughVID 9.62 vs. 9.79 yr. This is consistent with the breath-regression finding of Zhang et al. (2024), suggesting the advantage extends beyond breath audio.

3.3. Regression Head Comparison

MLP-small wins in 23 of 30 model × task combinations (Table 4), outperforming linear probing by up to 0.38 yr (HEAR on Coswara). Full MLP overfits on CIDRZ ($N_{\text{train}} = 669$): M2D+RESP degrades +0.53 yr vs. MLP-small, driven by a ∼22,000:1 parameter-to-sample ratio in the 3840-d hidden layer. On larger CoughVID ($N_{\text{train}} = 3050$), full MLP recovers and OPERA-GT achieves its best result (9.53 yr), revealing a dataset-size × head-capacity trade-off as a deployment guideline.

3.4. Cross-Dataset Generalisation

Table 5 reports cross-dataset age MAE for the best model per transfer direction. CoughVID → CIDRZ achieves a negative gap (−0.17 yr), showing that large-scale web-collected data can substitute for scarce clinical training data. The re-

Table 4. Full three-head comparison (MAE, 5-seed mean). Age/BMI in natural units; abnormality and TB in probability units $[0, 1]$. oCT/oCE/oGT = OPERA-CT/OPERA-CE/OPERA-GT; M2D = M2D+RESP. **Bold** = best head per model × task.

	Linear					MLP-small					MLP				
Task	oCT	oCE	oGT	HeAR	M2D	oCT	oCE	oGT	HeAR	M2D	oCT	oCE	oGT	HeAR	M2D
CIDRZ Age	**10.47**	10.55	**10.44**	10.58	10.63	10.52	**10.51**	10.49	**10.29**	**10.40**	10.57	10.70	10.54	10.39	10.93
CIDRZ BMI	3.64	3.62	**3.67**	3.64	3.68	**3.60**	**3.60**	**3.67**	**3.60**	**3.63**	**3.60**	**3.60**	**3.67**	3.61	3.91
CIDRZ Abn	**0.326**	**0.324**	0.317	0.331	0.316	0.327	0.325	**0.316**	**0.328**	0.320	0.331	0.329	**0.316**	0.332	**0.315**
CIDRZ TB	**0.189**	**0.190**	0.191	0.191	0.194	**0.189**	0.191	**0.190**	**0.188**	**0.192**	0.191	0.193	0.191	0.189	0.199
Coswara Age	10.25	10.46	10.25	9.50	9.98	10.25	**10.44**	10.16	**9.12**	**9.58**	**10.24**	10.49	**10.14**	9.26	9.90
CoughVID Age	9.81	**9.88**	9.71	9.78	9.86	**9.79**	**9.88**	9.62	**9.61**	**9.79**	**9.79**	9.89	**9.53**	9.67	9.95

Table 5. Cross-dataset age generalisation (MLP-small, best model per row). Gap = cross − within MAE of the same model. **Bold** = negative gap (cross-dataset outperforms in-domain).

Train → Test	Model	Cross	Within	Gap
CoughVID → CIDRZ	OPERA-CE	**10.34**	10.51	−0.17
Coswara → CIDRZ	OPERA-CE	10.54	10.51	+0.03
Coswara → CoughVID	OPERA-CT	10.42	9.79	+0.63
CoughVID → Coswara	HEAR	10.05	9.12	+0.94
CIDRZ → CoughVID	HEAR	10.54	9.61	+0.94
CIDRZ → Coswara	HEAR	11.55	9.12	+2.43

verse fails severely: CIDRZ → Coswara degrades +2.43 yr (+26.6%), indicating that small clinical populations do not generalise to large crowdsourced ones. HEAR leads in 3 of 6 directions, consistent with its broad health-audio pretraining corpus.

3.5. Low-Data Regime

Figure 2 shows CIDRZ age MAE as a function of training set size (MLP-small, 3 seeds). HEAR reaches within 0.02 yr of its $N = 669$ performance at $N = 50$ samples, and M2D+RESP shows similarly flat curves across all sizes, indicating that large-scale pretraining produces embeddings that require minimal labeled data for regression. OPERA models show substantially higher variance at $N = 50$ (std up to ±0.22 yr) and continue improving to $N = 400$, consistent with their smaller pretraining corpus.

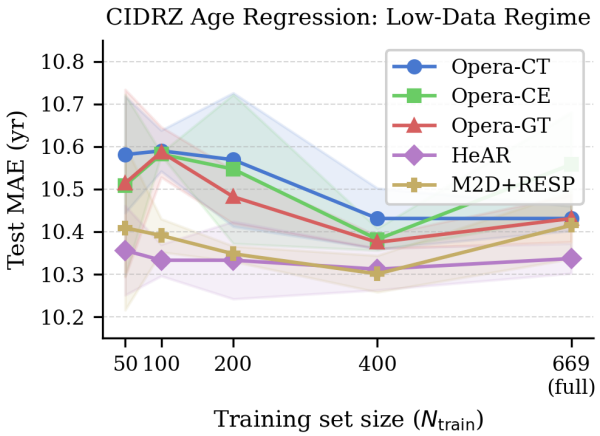

Figure 2. CIDRZ age MAE vs. training set size (MLP-small, 3-seed mean ± std). HEAR and M2D+RESP plateau by $N = 100$; OPERA models stabilise at $N = 400$.

4. Conclusion

We presented the multi-model, multi-target cough regression benchmark evaluating five respiratory acoustic FMs across six targets, three head architectures, and six cross-dataset transfer conditions. Four findings carry practical implications. First, MLP-small is the preferred head for frozen FM embeddings, outperforming linear probing in 23 of 30 cases while avoiding overfitting of full-width MLPs on small clinical datasets, though signal strength varies substantially across tasks (best/MAD ranging from 0.81 to 0.99), indicating that head choice matters most when the task has learnable structure. Second, generative pretraining (OPERA-GT) consistently outperforms contrastive pretraining (OPERA-CT) on age regression across all three datasets, extending a pattern observed for breath audio and consistent with the hypothesis that reconstruction-based objectives capture more physiological variation, though the mechanism warrants further study. Third, cross-dataset transfer is strongly asymmetric and large web-collected data generalises to small clinical populations with no performance loss while the reverse fails severely (+2.43 yr, 81× larger gap), likely reflecting the narrow demographic range of clinical cohorts relative to global crowdsourced corpora. Fourth, HEAR and M2D+RESP reach near-full performance at only $N = 50$ labeled samples while OPERA models require $N = 400$, indicating pretraining corpus diversity determines low-data regression performance over model architecture.

Limitations and future work. This study uses 2 s clips and evaluates cross-dataset transfer for age only, leaving BMI and clinical score generalisation across populations unexplored. FM fine-tuning was not evaluated and may further reduce the transfer gap on small clinical datasets. Future work should extend to additional targets, fine-tuning regimes, and attention-pooling heads, and the benchmark serves as a diagnostic for quantifying what physiological information respiratory FM embeddings encode.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Baur, S., Shor, J., Shams, B., Weng, W.-H., Sharma, A., Mortazavi, B., Nushi, B., and Venugopalan, S. HeAR: Health acoustic representations. *arXiv preprint arXiv:2403.02522*, 2024.

Bhattacharya, D., Sharma, N. K., Muguli, A., Kumar, P., Chetupalli, S. R., and Ganapathy, S. Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection. *Scientific Data*, 10(1):397, 2023.

Niizumi, D., Takeuchi, D., Yasuda, M., Nguyen, B. T., Ohishi, Y., and Harada, N. Towards pre-training an effective respiratory audio foundation model. In *Proc. Interspeech 2025*, 2025.

Orlandic, L., Teijeiro, T., and Atienza, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1): 156, 2021.

Rudraraju, G., Palreddy, S., Mamidgi, B., Sripada, N. R., Sai, Y. P., Vodnala, N. K., and Haranath, S. P. Cough sound analysis and objective correlation with spirometry and clinical diagnosis. *Informatics in Medicine Unlocked*, 19:100319, 2020.

Sharan, R. V., Xiong, H., and Berkovsky, S. Predicting spirometry readings using cough sound features and regression. *Physiological Measurement*, 39(9):095001, 2018.

World Health Organization. Global tuberculosis report 2021. Technical report, World Health Organization, Geneva, 2021.

Xu, W., Gu, L., Ma, L., Shen, X., and Rui, W. A forced cough sound based pulmonary function assessment method by using machine learning. *Frontiers in Public Health*, 10:1015876, 2022.

Zhang, Y., Xia, T., Han, J., Wu, Y., Rizos, G., Liu, Y., Mosuily, M., Chauhan, J., and Mascolo, C. Towards open respiratory acoustic foundation models: Pretraining and benchmarking. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.