nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00773-8

Audio-based AI classifiers show no evidence of improved COVID-19 screening over simple symptoms checkers

Received: 19 January 2023

Accepted: 19 November 2023

Published online: 7 February 2024

Check for updates

Harry Coppock[1,2,15], George Nicholson[1,3,4,15], Ivan Kiskin[1,3,5,15], Vasiliki Koutra[1,6], Kieran Baker[1,6], Jobie Budd[7,8], Richard Payne[9], Emma Karoune[1], David Hurley[9], Alexander Titcomb[9], Sabrina Egglestone[9], Ana Tendero Cañadas[9,10], Lorraine Butler[9], Radka Jersakova[1], Jonathon Mellor[9], Selina Patel[9,11], Tracey Thornley[12], Peter Diggle[13], Sylvia Richardson[1], Josef Packham[9], Björn W. Schuller[1,2,14], Davide Pigoli[1,6,16], Steven Gilmour[1,6,16], Stephen Roberts[1,3,16] & Chris Holmes[1,3,16] ✉

Recent work has reported that respiratory audio-trained AI classifiers can accurately predict SARS-CoV-2 infection status. However, it has not yet been determined whether such model performance is driven by latent audio biomarkers with true causal links to SARS-CoV-2 infection or by confounding effects, such as recruitment bias, present in observational studies. Here we undertake a large-scale study of audio-based AI classifiers as part of the UK government's pandemic response. We collect a dataset of audio recordings from 67,842 individuals, with linked metadata, of whom 23,514 had positive polymerase chain reaction tests for SARS-CoV-2. In an unadjusted analysis, similar to that in previous works, AI classifiers predict SARS-CoV-2 infection status with high accuracy (ROC–AUC = 0.846 [0.838–0.854]). However, after matching on measured confounders, such as self-reported symptoms, performance is much weaker (ROC–AUC = 0.619 [0.594–0.644]). Upon quantifying the utility of audio-based classifiers in practical settings, we find them to be outperformed by predictions on the basis of user-reported symptoms. We make best-practice recommendations for handling recruitment bias, and for assessing audio-based classifiers by their utility in relevant practical settings. Our work provides insights into the value of AI audio analysis and the importance of study design and treatment of confounders in AI-enabled diagnostics.

The coronavirus disease 2019 (COVID-19) pandemic has been estimated by the World Health Organization (WHO) to have caused 14.9 million excess deaths over the 2020–2021 period[1]. An accepted public health control measure for emerging infectious diseases is the isolation of infected individuals[2]. As COVID-19 transmission occurs in both symptomatic and asymptomatic cases[3], especially prior to nationwide vaccination deployment, a scalable and accurate test for the infection is crucial to avoid general population quarantine.

This has sparked an intense interest in AI-based classifiers that use respiratory audio data to classify severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection status (which we here refer to as COVID-19 status) via a digital screening tool that anyone with a

A full list of affiliations appears at the end of the paper. ✉e-mail: cholmes@stats.ox.ac.uk

smartphone or computer can use[4–18]. In our review, as of July 2022, we found 93 published papers that reported evidence for the potential of audio-based COVID-19 classification. Of these 93 papers, 75 report an area under the curve (AUC) (or F1) of over 0.75, whereas 44 report a performance of above 0.90. Extended Data Table 1 summarizes nine highly cited datasets and their corresponding classification performance.

Despite these encouraging results, concerns remain that the prediction models may not be transferable to real-world settings[11,15,18–21]. In some cases, data quality may be lowered by, for example, sampling biases, lack of verification of participants' COVID-19 status, a long delay between infection and audio recording, or small numbers of individuals who are SARS-CoV-2 reverse transcription polymerase chain reaction (PCR)-positive (COVID$^+$)[21]. Akin to findings in AI radiographic COVID-19 detection[22], concerns centre around whether the learnt audio features are unique audio biomarkers caused by COVID-19 in the infected individual, or are due to other confounding signals.

Here we analyse the largest PCR-validated dataset collected so far in the field of audio-based COVID-19 screening (ABCS). We design and specify an analysis plan in advance to investigate whether using audio-based classifiers can improve the accuracy of COVID-19 screening over using self-reported symptoms.

Our contribution is as follows:

- We collect a respiratory acoustic dataset of 67,842 individuals with linked PCR test outcomes, including 23,514 who tested positive for COVID-19. This is, to the best of our knowledge, the largest PCR-validated dataset collected of its kind so far[23].
- We fit a range of AI classifiers and observe strong COVID-19 predictive performance (receiver operating charateristic area under the curve (ROC–AUC) = 0.85), as has been reported in past studies, for example refs. 4–18; however, when controlling for measured confounders by matching, only a small amount of residual predictive variation remains (ROC–AUC = 0.62), some of which we attribute to unmeasured confounders.
- We find the COVID-19 predictive performance and practical utility of audio-based AI classifiers—as applied in simulated realistic settings—to be no better than classification on the basis of self-reported symptoms; we replicate this finding by fitting our classifiers in an external dataset.
- These results suggest that audio-based classifiers learn to predict COVID-19 via self-reported symptoms and potentially other confounders. Study recruitment on the basis of self-screened symptoms seems to be an important driver of this effect.
- We provide best-practice recommendations on how to address this problem in future studies.
- Our dataset and code-base is publicly available to enable reproducibility of results and to encourage further research into respiratory audio analysis and bias mitigation in high-dimensional, over-parameterized settings[23].

Our work is timely in highlighting the need for careful construction of machine learning evaluation procedures, aimed at yielding representative performance metrics. The important lessons from this case study on the effects of confounding extend across many applications in AI—where biases are often hard to spot and difficult to control for.

Results

Study design

This study invited volunteers from the Real-time Assessment of Community Transmission (REACT) programme and the National Health Service (NHS) Test-and-Trace (T+T) service to participate between March 2021 and March 2022 on an opt-in basis. Volunteers were directed to the 'Speak up and help beat coronavirus' web page[24], where they were instructed to provide audio recordings of four respiratory audio modalities. Demographic and health metadata, along with a validated PCR test result, were transferred from existing T+T/REACT records. Further audio-specific metadata were produced from the audio files after collection. The final dataset comprised 23,514 COVID$^+$ and 44,328 SARS-CoV-2 PCR-negative (COVID$^-$) individuals. Figure 1 summarizes the dataset (a more detailed description of which is provided in Methods) and a full presentation can be found in the accompanying dataset paper[23].

Defining the acoustic target for COVID-19 screening

If a practically effective acoustic signal were to exist in SARS-CoV-2-infected individuals' respiratory sounds, we propose that it would have the following properties:

P1: Caused by COVID-19. COVID-19 is well known to cause symptoms (such as a new continuous cough) that can be readily self-screened by individuals in the general population. The acoustic target would likewise be linked causally to COVID-19 and would therefore be more likely to generalize to other contexts and populations than non-causal associations.

P2: Not self-identifiable. The acoustic target would not directly represent self-identifiable symptoms that can be self-identified effectively by individuals in the general population. This is because: (1) it is more straightforward to measure self-identifiable symptoms directly using a questionnaire, rather than measuring them indirectly via audio; and (2) as we explain below, self-identifiable symptoms can affect enrolment and may therefore be strongly non-causally associated with COVID-19 in enrolled subpopulations.

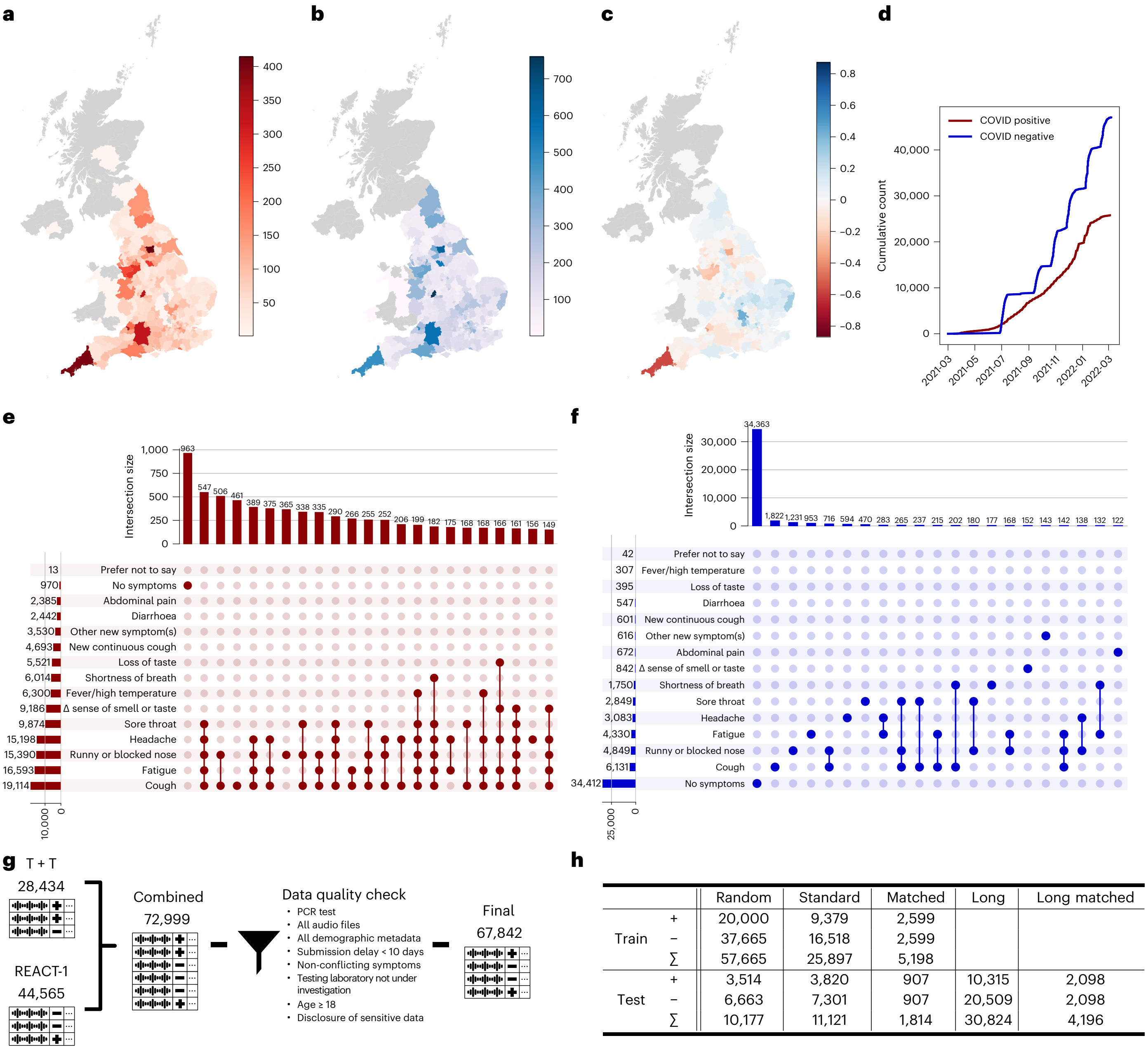
P3: Enables high-utility COVID-19 screening. For an audio-based classifier to perform strongly in practical settings, it should possess high sensitivity and specificity, corresponding to an acoustic signal that would be detectable in high and low proportions of individuals who are COVID$^+$ and COVID$^-$, respectively. We formalize the mathematical relationship linking expected utility, sensitivity and specificity in equation (1) (see Methods).

Characterizing and controlling recruitment bias

In audio-based COVID-19 classification, results can be highly sensitive to the characteristics of the enrolled population. Our study's recruitment protocol is subject to enrolment bias because the vast majority of individuals in pillar 2 of the UK government's NHS T+T programme[25] were PCR tested as a direct consequence of reporting symptoms (see Methods). Figure 1e,f display our participants' symptom profiles, stratified by COVID-19 infection status. Figure 2a presents the joint distribution of COVID-19 status and binary symptoms status as 'symptoms-based enrolment', in contrast to Fig. 2b, which presents 'general population enrolment' on the basis of random sampling from a general population with 2% COVID$^+$ prevalence. Note that the dependence between binary symptoms status and COVID-19 is stronger under symptoms-based enrolment (population correlation coefficient $\rho = 0.66$) than general population enrolment ($\rho = 0.15$).

We will discuss three simplified recruitment processes to illustrate the effects of different types of enrolment protocol. These are illustrated in Fig. 3 using the probabilistic framework of directed and undirected graphical models (a good introduction to which can be found in chapters 10 and 19 of ref. 26). As defined above, our goal is to train a classifier capable of predicting COVID-19 via its latent acoustic signature. We explain below how this requires the classifier predictions to be conditionally dependent on the latent COVID signature given self-reported symptoms, denoted by the red edge in Fig. 3ciii under a matched recruitment protocol; however the corresponding edges are missing under the other recruitment protocols in Fig. 3aiii and Fig. 3biii.

First consider our simplified causal model of symptoms-based recruitment (Fig. 3ai). Enrolment is jointly influenced by COVID-19 status, self-reported symptoms and factors such as age and gender (Extended Data Fig. 1 shows a detailed Bayesian knowledge graph of the



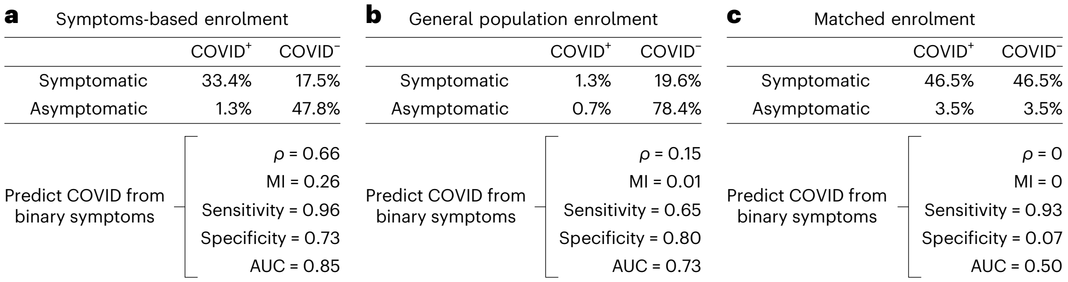
		Random	Standard	Matched	Long	Long matched
Train	+	20,000	9,379	2,599		
	−	37,665	16,518	2,599		
	∑	57,665	25,897	5,198		
Test	+	3,514	3,820	907	10,315	2,098
	−	6,663	7,301	907	20,509	2,098
	∑	10,177	11,121	1,814	30,824	4,196

Fig. 1 | Demographic statistics of collected dataset. a,**b**, Geographical locations of COVID positive (**a**) and negative (**b**) PCR-confirmed participants. Colour bar units are individual participant count. **c**, $\%\ \frac{100\times\#\text{negative participants at location}}{\#\text{negative participants in total}} - \frac{100\times\#\text{positive participants at location}}{\#\text{positive participants in total}}$ Colour bar units are the difference between percentage points. **d**, Cumulative count of the number of participants partaking in the study. **e**,**f**, The 21 most common combinations of symptoms for COVID positive (**e**) and negative (**f**) participants, ordered along the x-axis by total number of participants displaying that particular combination of symptoms. Symptoms are ordered along the y-axis according to total number of participants displaying at least that symptom at the time of recording. **g**, Schematic detailing the two recruitment sources for the study and the filtration steps applied to yield the final dataset. **h**, Dataset splits in participant numbers.

recruitment process). Collecting data only from enrolled individuals is, in effect, conditioning on $e_i = 1$ at the enrolment node in Fig. 3ai. As the enrolment node has directed edges incoming from both COVID-19 status and self-reported symptoms (that is, it is a collider node), conditioning on it induces a non-causal dependence between its parent nodes (in addition to the causal dependence of symptoms on COVID-19 status). Figure 3aii displays the moralized undirected graph implied by Fig. 3ai, conditional on enrolment, with the strong COVID-19-to-symptoms dependence represented by a thick line illustratively labelled $\rho = 0.66$ with reference to Fig. 2a. By contrast, Fig. 3bi is conditional on random enrolment and does not introduce any additional non-causal association between COVID-19 status and self-reported symptoms.

If a study's enrolment bias is unaddressed and shared across both training and held-out test sets, a classifier that seems to perform well may not generalize to other datasets[20,21]. This is due to two effects: first, the classifier may learn to predict using confounding variables that are not causally related to COVID-19 but are associated due to their influence on enrolment (for example, gender, age or symptoms unrelated to COVID-19); second, even symptoms that are truly causally related to COVID-19, such as a new continuous cough, may exhibit inflated association with COVID-19 in the enrolled cohort due to their influence on enrolment (illustrated by the thick edges labelled $\rho = 0.66$ in Fig. 3aii,aiii).

As well as leading to poor generalizability, audible characteristics that are non-causally but strongly associated with COVID-19



a Symptoms-based enrolment

	COVID$^+$	COVID$^-$
Symptomatic	33.4%	17.5%
Asymptomatic	1.3%	47.8%

Predict COVID from binary symptoms

ρ = 0.66
MI = 0.26
Sensitivity = 0.96
Specificity = 0.73
AUC = 0.85

b General population enrolment

	COVID$^+$	COVID$^-$
Symptomatic	1.3%	19.6%
Asymptomatic	0.7%	78.4%

Predict COVID from binary symptoms

ρ = 0.15
MI = 0.01
Sensitivity = 0.65
Specificity = 0.80
AUC = 0.73

c Matched enrolment

	COVID$^+$	COVID$^-$
Symptomatic	46.5%	46.5%
Asymptomatic	3.5%	3.5%

Predict COVID from binary symptoms

ρ = 0
MI = 0
Sensitivity = 0.93
Specificity = 0.07
AUC = 0.50

Fig. 2 | Illustrative tables relating symptoms status with COVID-19 status. **a**, Symptoms-based enrolment, where individuals who are COVID$^+$ are preferentially recruited on the basis of symptoms (percentages are calculated from the entire sample of individuals recruited into this study). **b**, General population enrolment on the basis of random sampling from an illustrative general population with a COVID-19 prevalence of 2%, where symptomatic individuals make up 20% and 65% of COVID$^-$ and COVID$^+$ subpopulations, respectively. **c**, Matched enrolment, where the number of individuals who are COVID$^-$ and COVID$^+$ is the same for each particular symptoms profile within the symptomatic and asymptomatic subgroups (percentages shown are for the matched test set in the current study). For each type of enrolment, the diagnostic accuracies of the resulting symptoms-only COVID-19 classifier are shown below the table: ρ, mutual information (MI), sensitivity, specificity and AUC.

can obscure any COVID-19 acoustic signature that may exist. This is illustrated in Fig. 3aiii, where the association between classifier prediction and SARS-CoV-2 status is mediated by symptoms instead of via a targeted latent COVID acoustic signature (that is, there are no edges corresponding to the red edge seen in Fig. 3ciii). Even in the case of randomized enrolment from the general population, a classifier may learn to predict SARS-CoV-2 status via self-reported symptoms, as opposed to via a latent COVID-19 acoustic signature, as illustrated in Fig. 3biii (again, a lack of red edge indicates that classifier predictions are conditionally independent of latent acoustic signature given self-reported symptoms).

Here, our goal is to build a classifier whose association with COVID-19 is mediated by an acoustic signature with the three properties defined above. We use the established epidemiological methodology known as matching[27], whereby study enrolment balances the number of COVID$^+$ and COVID$^-$ participants having each combination of potentially audible measured confounding variables. This has the effect of inducing independence between COVID-19 and these confounders in the matched population, as shown in Fig. 3cii. The classifier is then constrained to predict COVID-19 status either via the latent COVID-19 acoustic signature (via the red edge in Fig. 3ciii), or via unmeasured confounders.

Primary analyses

Pre-specified analysis plan. We designed and fixed a pre-specified analysis plan to increase the replicability of conclusions[28]. As part of this advance planning, we detailed the analyses to be conducted and generated the test/validate/train data splits through subsampling of the full dataset. The design of these splits is detailed in the Methods, with sample sizes listed in Fig. 1h.

Audio-based COVID-19 prediction performance. Table 1 presents our study's COVID-19 prediction performance across nine train/validate/test splits, four modalities and three models: Self-Supervised Audio Spectrogram Transformer (SSAST), Bayesian neural networks (BNNs) and an openSMILE–support vector machine (SVM). The SSAST and BNN classifiers consistently outperform the baseline SVM, and the best prediction is achieved with the sentence modality. Reported results are for the SSAST performance on the sentence modality, unless stated otherwise. Under the randomized data split, the SSAST classifier achieves a high COVID-19 predictive accuracy of ROC–AUC = 0.846 [0.838–0.854]. We hypothesize that this strong predictive accuracy is mainly attributable to enrolment on the basis of self-reported symptoms, and explore this further in confirmatory analyses below.

When we control for enrolment bias by matching on age, gender and self-reported symptoms, predictive accuracy drops to a consistently low level of ROC–AUC = 0.619 [0.594–0.644] in the matched test set, and ROC–AUC = 0.621 [0.605–0.637] in the longitudinal matched test set—that is, a temporally out-of-distribution test set consisting of only submissions after 29 November 2021 (both trained on the standard training set). When training instead on our matched training set, we see a minor improvement in the matched test set (ROC–AUC = 0.635 [0.610–0.660]), and, by contrast, a slight decrease in prediction accuracy in the longitudinal matched test set (ROC–AUC = 0.604 [0.588–0.620]). Figure 4 illustrates these different experimental settings and the corresponding classification performance. A cluster analysis is also performed on the SSAST learnt representations (detailed in Supplementary Note 2) visually demonstrating the effect of decoupling measured confounders and COVID-19 status. To explore whether classifier performance might be higher in some matched groups than in others, we calculated ROC–AUC within matched strata (Extended Data Fig. 2), observing the estimates and confidence intervals to be consistent with a homogeneously low predictive score of ROC–AUC = 0.62 across strata.

Confirmatory analyses and validation

The additional predictive value of ABCS. Audio-based classifiers can be useful in practice if they deliver improved performance relative to classifiers that are based on self-identifiable symptoms. Moreover, it is beneficial to assess the performance of ABCS classifiers in test sets reflecting the application of the testing protocol in a real-life setting. Here we generate a general population test set, through balanced subsampling, without replacement from our combined standard and longitudinal test sets, to capture the age/gender/symptoms/COVID-19 profile of the general population during the pandemic. Specifically, the proportion of symptomatic individuals is set to 65% in the COVID$^+$ subgroup[29], compared with a setting of either (10%, 20%, 30%) symptomatic individuals in the COVID$^-$ subgroup; the age distribution is constrained to be the same in both COVID$^+$ and COVID$^-$ subgroups; and with males/females balanced in a 1:1 ratio in each COVID$^+$/COVID$^-$ subgroup. We benchmark the COVID-19 predictive performance of the audio-based SSAST classifier against the performance attainable through random forest (RF) classifiers trained on self-identifiable symptoms and demographic data (a 'symptoms' RF classifier). In the benchmarking we also include an RF classifier, which takes as inputs the audio-based SSAST probabilistic outputs alongside self-identifiable symptoms and demographic data ('symptoms + audio' RF classifier). Training for all three classifiers is performed in our standard training set. The resulting ROC curves are shown in Fig. 5a–c. Focusing on the general population with 20% of COVID$^-$ symptomatic in Fig. 5b, the combined symptoms + audio RF classifier (ROC–AUC = 0.787 [0.772–0.801], 95% DeLong CI) offers a significant ($P = 9.7 \times 10^{-11}$, DeLong test) but small increase in predictive accuracy over the symptoms RF classifier (ROC–AUC = 0.757 [0.743–0.771], 95% DeLong CI), which in turn yields a less significant increase

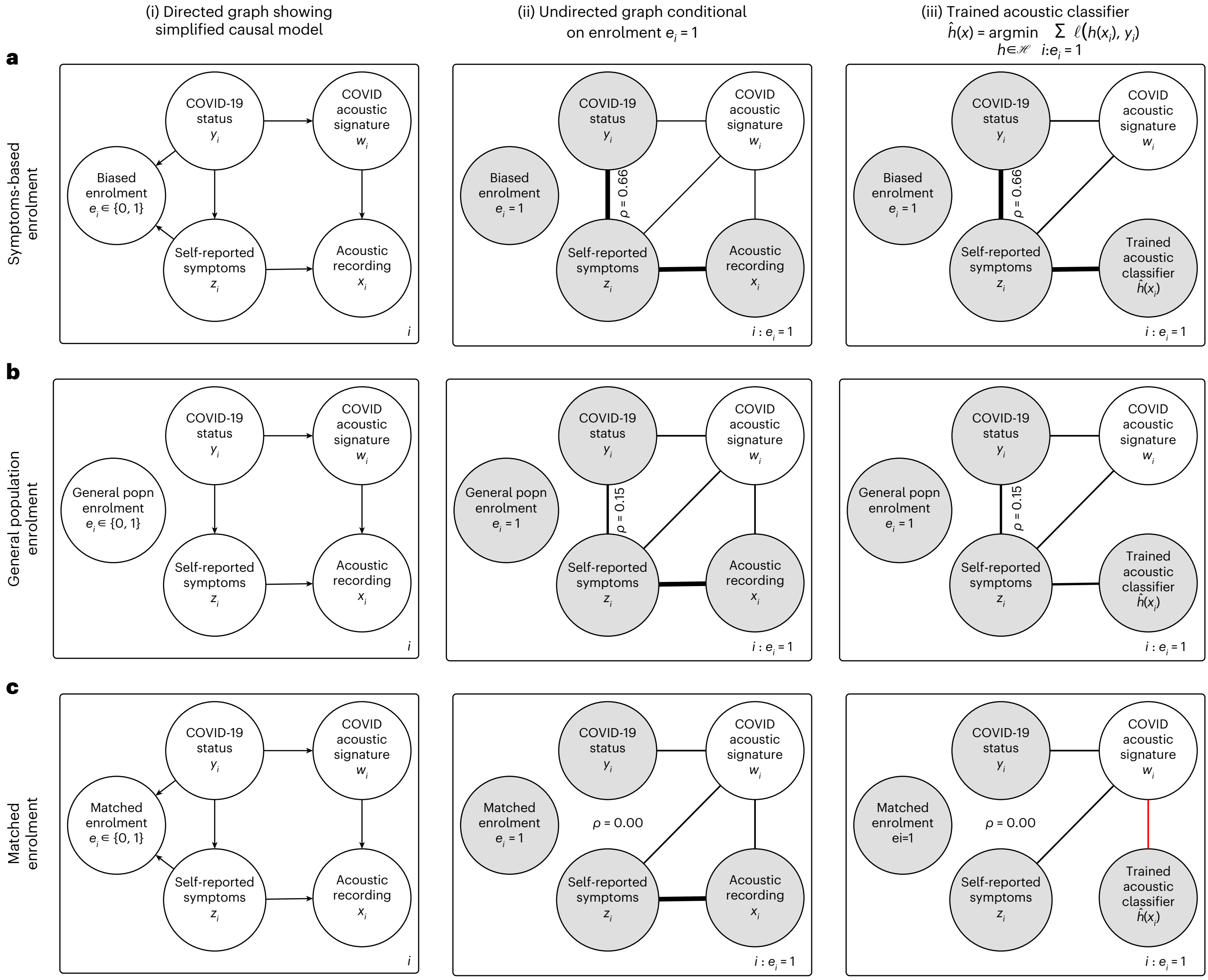

(i) Directed graph showing simplified causal model
(ii) Undirected graph conditional on enrolment $e_i = 1$
(iii) Trained acoustic classifier $\hat{h}(x) = \text{argmin}_{h \in \mathscr{H}} \sum_{i:e_i=1} \ell(h(x_i), y_i)$
a
Symptoms-based enrolment
COVID-19 status y_i
COVID acoustic signature w_i
Biased enrolment $e_i \in \{0, 1\}$
Self-reported symptoms z_i
Acoustic recording x_i
i
Biased enrolment $e_i = 1$
$\rho = 0.66$
$i : e_i = 1$
Trained acoustic classifier $\hat{h}(x_i)$
b
General population enrolment
General popn enrolment $e_i \in \{0, 1\}$
General popn enrolment $e_i = 1$
$\rho = 0.15$
c
Matched enrolment
Matched enrolment $e_i \in \{0, 1\}$
Matched enrolment $e_i = 1$
Matched enrolment ei=1
$\rho = 0.00$

Fig. 3 | Study enrolment influences capability of classifiers to learn a latent COVID acoustic signature. a–**c**, Our goal is to train a classifier whose predictions are conditionally dependent on the latent COVID acoustic signature given self-reported symptoms, as denoted by the red edge in **ciii** under matched enrolment. This edge is not present in **aiii** or **biii** because the classifier's predictive ability is mediated by self-reported symptoms under symptoms-based and general population enrolment. Rows **a**–**c** present three different enrolment protocols, whereas columns (i)–(iii) show different types of graph; in the undirected graphs, shaded nodes are observed variables, while the edge thickness is used to illustrate the approximate strength of conditional dependence. **a**, Symptoms-based enrolment enforces a supervised sampling regime in which individuals who are COVID$^+$ are preferentially recruited on the basis of symptoms shown (for example, Fig. 2a). **b**, Randomized enrolment performs a random sampling of individuals from the general population. **c**, Matched enrolment balances the number of COVID$^+$ and COVID$^-$ individuals that share each particular symptoms profile; (i) Bayesian knowledge graphs displaying a simplified causal model; (ii) undirected conditional independence graphs implied by the directed graphs in (i) when we condition on enrolment ($e_i = 1$); (iii) undirected conditional independence graphs, as in (ii), but now showing the trained acoustic classifier $\hat{h}(x_i)$ (trained to predict y_i based on input x_i) in lieu of the acoustic recording data x_i.

in ROC–AUC compared with the audio-only classifier ($P = 0.0033$) (ROC–AUC = 0.733 [0.717–0.748], 95% DeLong CI).

We replicate these findings using an external dataset[18], observing qualitatively similar results with a symptoms classifier (ROC–AUC = 0.79 [0.71–0.87]) outperforming an SSAST audio-based classifier (ROC–AUC = 0.68 [0.59–0.77]) in a general population test set simulated from the external test set[18], in which 20% of the individuals who are COVID$^-$ are symptomatic. We observe similar results when comparing a symptoms classifier with our SSAST and Han and colleagues' convolutional neural network (CNN) on the external test set directly[18]: symptoms classifier ROC–AUC = 0.81 [0.76–0.86]; SSAST audio-based classifier ROC–AUC = 0.68 [0.62–0.74]; CNN audio-based classifier[18] ROC–AUC = 0.66 [0.60–0.71] (see also Extended Data Fig. 3). The reported results for our SSAST and Han and colleagues' CNN are for the 'cough' modality; however, we see similar results for both 'breathing' and 'voice'.

Translating prediction accuracy into utility. To characterize the practical benefit of ABCS in any particular testing setting, we can specify the utility $u_{\hat{y},y}$ of predicting $\hat{y} \in \{0, 1\}$ for a random individual, in the targeted testing population, whose true COVID status is $y \in \{0, 1\}$, and calculate the per-test expected utility (EU) as

$$\begin{aligned}\text{EU} &\equiv \mathbb{E}[\text{utility}|\boldsymbol{u}, \pi, \text{sensitivity}, \text{specificity}]\\ &= \pi\left[(u_{1,1} - u_{0,1}) \times \text{sensitivity} + u_{0,1}\right] \qquad (1)\\ &\quad +(1-\pi)\left[(u_{0,0} - u_{1,0}) \times \text{specificity} + u_{1,0}\right],\end{aligned}$$

Table 1 | Results detailing the performance of the SVM, SSAST and BNN models on the nine evaluation tasks for each of the four audio modalities: sentence, three coughs, cough and exhalation

Train			Standard (9,379$^+$16,518$^-$)				Match (2,599$^+$2,599$^-$)				Random (20,000$^+$ 37,665$^-$)
Test			**Standard (3,820$^+$7,301$^-$)**	**Match (907$^+$907$^-$)**	**Long (10,315$^+$20,509$^-$)**	**Long match (2,098$^+$2,098$^-$)**	**Standard (3,820$^+$7,301$^-$)**	**Match (907$^+$907$^-$)**	**Long (10,315$^+$ 20,509$^-$)**	**Long match (2,098$^+$ 2,098$^-$)**	**Random (3,514$^+$ 6,663$^-$)**
Sentence	SVM	UAR	0.669	0.566	0.699	0.570	0.658	0.567	0.646	0.579	0.721
		ROC	0.732	0.596	0.766	0.591	0.714	0.600	0.693	0.597	0.796
		PR	0.578	0.574	0.625	0.580	0.553	0.583	0.515	0.576	0.686
	SSAST	UAR	**0.733**	**0.594**	**0.739**	**0.583**	0.692	0.602	0.666	0.572	**0.763**
		ROC	**0.800**	0.619	**0.818**	**0.621**	0.760	**0.635**	0.732	0.604	**0.846**
		PR	**0.684**	0.594	**0.715**	**0.594**	0.631	0.626	0.590	0.579	**0.774**
	BNN	UAR	0.685	0.586	0.702	0.566	0.703	**0.604**	0.687	0.581	0.702
		ROC	0.776	**0.623**	0.804	0.614	0.767	0.634	0.749	0.610	0.834
		PR	0.645	**0.613**	0.689	0.593	0.634	**0.629**	0.619	0.593	0.752
Three coughs	SVM	UAR	0.669	0.555	0.694	0.541	0.635	0.539	0.639	0.550	0.713
		ROC	0.727	0.568	0.759	0.558	0.684	0.560	0.688	0.568	0.782
		PR	0.570	0.550	0.605	0.538	0.523	0.553	0.510	0.546	0.647
	SSAST	UAR	0.681	0.555	0.696	0.551	0.652	0.546	0.662	0.555	0.725
		ROC	0.750	0.577	0.781	0.569	0.714	0.571	0.723	0.568	0.809
		PR	0.607	0.553	0.648	0.552	0.563	0.557	0.561	0.557	0.701
	BNN	UAR	0.678	0.558	0.696	0.551	0.657	0.558	0.660	0.535	0.716
		ROC	0.751	0.578	0.786	0.578	0.713	0.578	0.720	0.558	0.807
		PR	0.601	0.550	0.647	0.556	0.551	0.554	0.563	0.551	0.691
Cough	SVM	UAR	0.648	0.536	0.685	0.540	0.633	0.541	0.638	0.538	0.695
		ROC	0.712	0.544	0.748	0.550	0.687	0.559	0.692	0.559	0.763
		PR	0.559	0.526	0.594	0.535	0.533	0.550	0.521	0.545	0.625
	SSAST	UAR	0.681	0.545	0.690	0.541	0.638	0.528	0.640	0.543	0.702
		ROC	0.742	0.561	0.768	0.559	0.692	0.552	0.692	0.560	0.790
		PR	0.603	0.540	0.631	0.548	0.535	0.545	0.532	0.550	0.675
	BNN	UAR	0.647	0.540	0.661	0.534	0.618	0.532	0.638	0.541	0.672
		ROC	0.732	0.570	0.765	0.563	0.682	0.542	0.698	0.556	0.786
		PR	0.581	0.556	0.621	0.549	0.511	0.526	0.522	0.541	0.678
Exhalation	SVM	UAR	0.600	0.523	0.639	0.544	0.587	0.528	0.585	0.529	0.653
		ROC	0.646	0.555	0.690	0.559	0.618	0.541	0.621	0.550	0.712
		PR	0.477	0.560	0.513	0.547	0.444	0.536	0.431	0.543	0.566
	SSAST	UAR	0.649	0.553	0.663	0.558	0.593	0.531	0.588	0.531	0.660
		ROC	0.701	0.581	0.725	0.580	0.653	0.552	0.644	0.556	0.750
		PR	0.563	0.578	0.575	0.561	0.496	0.548	0.473	0.549	0.634
	BNN	UAR	0.576	0.529	0.581	0.526	0.603	0.525	0.601	0.541	0.608
		ROC	0.683	0.569	0.722	0.578	0.679	0.570	0.675	0.567	0.744
		PR	0.539	0.581	0.573	0.563	0.519	0.573	0.507	0.551	0.620

The metrics corresponding to the highest performance for each of the 18 (evaluation procedure, test set) pairs (that is, for each pair in {UAR, ROC, PR}×{standard, match, long, long match, random}) across all modalities and models, are bolded. Each training and test set is shown with the corresponding support of individuals who are COVID$^+$ and COVID$^-$. ROC, ROC–AUC; PR, PR–AUC; UAR, unweighted average recall.

where π is the COVID-19 prevalence in the tested population (equation (1) is derived in the Methods). The EU is increasing in both sensitivity and specificity, with their relative weights depending on prevalence π and utility $\mathbf{u}$. Note that it is not only π and $\mathbf{u}$ that are context-dependent: the sensitivity and specificity of any particular COVID-19-classifier depends on the characteristics of the targeted testing population, as illustrated by the effects of variation in the proportion of COVID$^-$ individuals that are symptomatic across Fig. 5a–c.

Consider the following illustrative utility function, measured in units of the number of infections prevented:

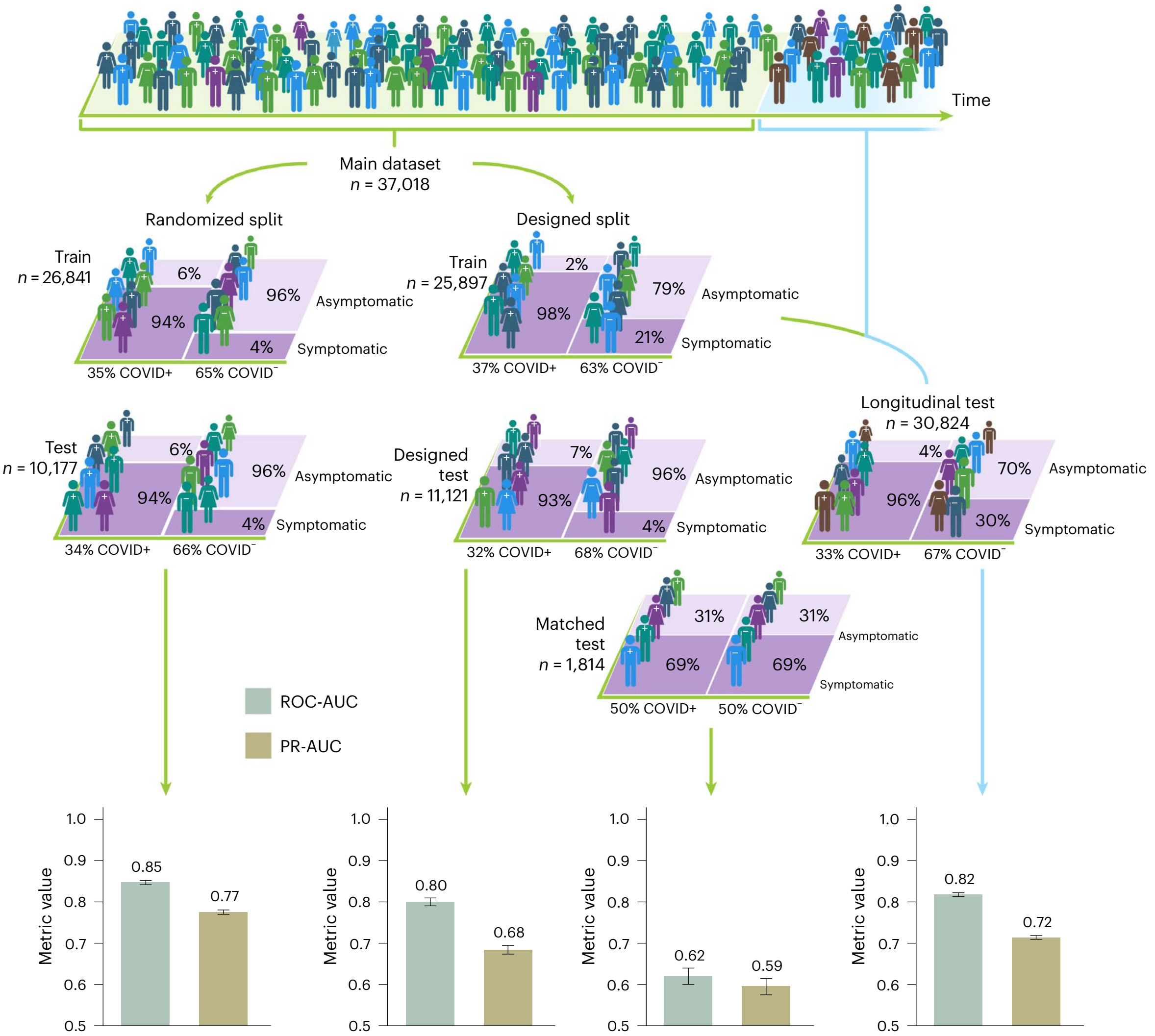


Fig. 4 | Schematic demonstrating the importance of ascertainment bias adjustment in the context of reporting COVID-19 detection from patient respiratory audio performance. Human figures represent study participants and their corresponding COVID-19 infection status, with the different colours portraying different demographic or symptomatic features. When participants are randomly split into training and test sets, the randomized split models perform well at COVID-19 detection, achieving AUCs in excess of 0.8; however, matched test set performance is seen to drop to estimated AUC between 0.60 and 0.65, with an AUC of 0.5 representing random classification. Inflated classification performance is also seen in engineered out of distribution test sets such as: the designed test set, in which a select set of demographic groups appear solely in the testing set, and the longitudinal test set, in which there is no overlap in the time of submission between train and test instances. The 95% confidence intervals calculated via the normal approximation method are shown, along with the corresponding n numbers of the train and test sets. Figure 4 created with Biorender.com.

$$
\begin{aligned}
u_{1,1} &= R_t - \varepsilon \\
&\quad \left[\text{True positive result for COVID}^{+},\right. \\
&\quad \left. R_t \text{ infections prevented on average}\right] \\
u_{1,0} &= -\varepsilon \\
&\quad \left[\text{False positive for COVID}^{-},\right. \\
&\quad \left. -\varepsilon \text{ is negative impact of self-isolation}\right] \\
u_{0,0} &= 0 \\
&\quad \left[\text{True negative for COVID}^{-}\right] \\
u_{0,1} &= -\delta \\
&\quad \left[\text{False negative for COVID}^{+},\right. \\
&\quad \left. \text{causing } \delta \text{ additional infections on average}\right].
\end{aligned}
$$

There are three specified parameters in the above: (i) the number of cases prevented by intervention on a single individual is specified as the effective reproduction number, $R_t \geq 0$, that is, the average number of infections that person would cause under no intervention, assuming that all individuals with a positive result follow self-isolation guidance and cause no transmission; (ii) $\varepsilon \geq 0$ measures the cost of intervention (for example, the negative impact on health or education resulting from self-isolation); and (iii) $\delta \geq 0$ is the expected number of additional infections caused by a false-negative result (for example, due to reduced caution and increased social mixing following a negative result).

Figure 5d–f shows maximum EU, as a function of prevalence, for settings of $R_t \in \{1, 1.5\}$ and $\varepsilon \in \{0.02, 0.2\}$ and with $\delta = 0$ (corresponding results for $\delta = 0.25$ are shown in Extended Data Fig. 4). The maximization is performed point-wise with respect to sensitivity and specificity across the corresponding ROC curves in Fig. 5a–c. The utility of all classifiers decreases as the percentage of COVID⁻ symptomatics increases from 10% to 30% in Fig. 5d–f, with the intuition being that it is more

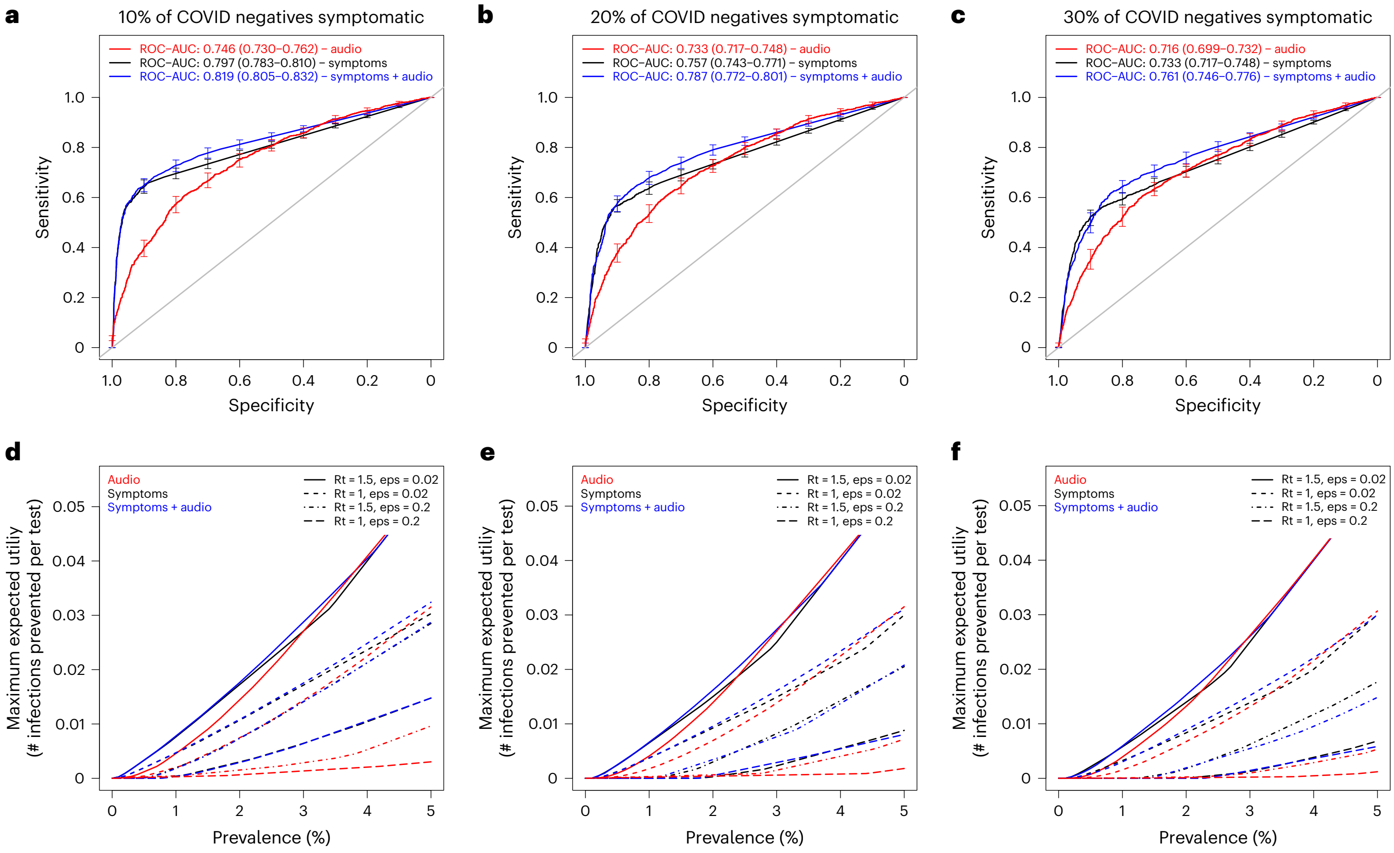


Fig. 5 | Comparison of sensitivity, specificity, and utility across audio-based and symptoms-based classifiers, as applied in simulated general populations. a–c, The percentage of COVID$^-$ individuals who are symptomatic in the general population varies between 10% and 30% across the three columns of panels (labelled at the top). Comparison of ROC curves between the audio, symptoms, and symptoms + audio classifiers; curves show sensitivity as a function of specificity, with error bars denoting 95% confidence intervals for sensitivity, where confidence intervals are calculated using the pROC::ci.se() R function and are based on a sample size of 2,000 stratified bootstrap replicates. The legends show the curve colour for each classifier alongside ROC–AUC estimates and 95% DeLong confidence intervals. **d–f**, Comparison of maximum expected utility across classifiers. Four different utility functions are included, as detailed in the top-right legend (utility function parameters R_t, ε and δ are defined in the 'Results' section; $\delta = 0$ in this figure). Curves are coloured to indicate audio, symptoms, or symptoms + audio classifiers, as detailed in the top-left legend.

difficult to distinguish COVID$^-$ from the 65% symptomatic COVID$^+$ population. When we compare the symptoms + audio RF classifier with the symptoms RF classifier, neither is generally optimal, with each classifier showing greater EU than the others for some values of (π, R_t, ε).

Exploratory approaches to identify the influence of unmeasured confounders. We explore whether the residual COVID-19 prediction performance of ROC–AUC = 0.62 in the matched test set is truly attributable to the targeted acoustic signature, or whether it stems from unmeasured confounders such as the audible recording environment, or unreported symptoms. We describe the two complementary exploratory methods in greater detail in Supplementary Note 1.

Method 1 investigates how much of the residual predictive variation persists when we map all matched test set samples to the first k-principal components of the COVID$^-$ samples. We train a classifier on COVID-19 detection in this k-dimensional space and hypothesize that, below a threshold value for $k \le \tau$, correct classification is due to confounding in the signal. The value of τ is determined by running a calibration experiment and is set to 14 for the sentence modality. By removing these correctly classified cases to form a curated matched test set, we see a drop in SSAST performance to UAR = 0.51 (the results from this experiment can be found in Extended Data Fig. 5).

Method 2 examines how much residual predictive variation persists when, using a robust distance metric in openSMILE space, we map each COVID$^+$ participant to their nearest COVID$^-$neighbour. Predictive variation that persists in the space spanned by individuals who are COVID$^-$ is then attributed to unmeasured confounders. After the COVID$^+$-to-COVID$^-$ nearest-neighbour mapping, the SVM matched test set ROC–AUC drops from 0.60 to 0.55. We interpret this persistent component of predictive variation after the mapping to COVID$^-$ individuals (that is, ROC–AUC drops only to 0.55 as opposed to 0.50) as pointing to some degree of unmeasured confounding contributing to the score of ROC–AUC = 0.60 in the matched test set.

Discussion

COVID-19 is well known to be causally related to particular self-identifiable symptoms such as a new continuous cough. This has allowed such symptoms to be used by governments during the pandemic as a basis for population intervention to control disease spread (for example, as a triage tool for individuals) via self-screening and without recourse to audio recording. It is therefore desirable to develop audio-based classifiers that can augment and complement the information provided by self-identifiable COVID-19-specific symptoms, that is, to learn clinically valuable latent acoustic signatures caused by COVID-19.

Problematically, enrolment on the basis of symptoms has the potential not only to artificially inflate the association between COVID-19 and its particular symptoms, but also to introduce association between COVID-19 and symptoms that are not COVID-19 specific. Furthermore, enrolment on the basis of other characteristics such as gender and age may also introduce non-causal COVID-19–gender or COVID-19–age associations in the enrolled subpopulation, possibly interacting with symptoms.

Under such recruitment bias, classifiers trained to predict COVID-19 in enrolled subpopulations may learn to predict self-identifiable COVID-19-specific symptoms, thereby providing no further utility beyond a classifier trained directly on those self-screened symptoms. It is worth noting that there are exceptions to this statement, including situations in which passive monitoring is necessary or when individuals have reasons to provide incorrect information. More concerningly, the classifier may learn to predict age/gender/non-COVID-19-specific symptoms as proxies for COVID-19 in the enrolled subpopulation, in which case its performance will not generalize to subpopulations unaffected by the same recruitment bias.

Han et al.[18] examine several aspects of recruitment bias in ABCS, simulating the effects of biases introduced by age, gender, language, and by the same individuals appearing in both the train and test sets. Although their dataset is approximately balanced with respect to age and gender (across COVID$^+$ and COVID$^-$ subgroups), it is imbalanced with respect to self-reported symptoms (84% of their COVID$^+$ subgroup is symptomatic, compared with 49% of the COVID$^-$subgroup; ref. 18, Fig. 1e). Such imbalance with respect to symptoms is also present in our study (prior to matching) and other studies for which data, including symptoms, are available (see Extended Data Fig. 6).

We make the following recommendations with the aim of clarifying the effects of recruitment bias and mediation by self-identifiable symptoms in future studies:

1. **Collect and disseminate metadata**. Repositories of audio samples should include details of the study recruitment criteria and relevant metadata (for example, gender, age, symptoms, location, time since COVID-19 test) so that data can be filtered for quality and for relevance to hypothesis, and bias from measured confounders can be characterized and controlled if necessary.
2. **Characterize and control recruitment bias**. Analyse data using methods that acknowledge and control for the effects of recruitment bias. We approached this by matching on measured confounders in our test and/or training sets.
3. **Design studies with bias control in mind**. Matching leads to reduced sample size when performed post-recruitment, so it can be beneficial to design observational studies that recruit participants to maximize the potential for matching on measured confounders in the enrolled cohort.
4. **Focus on the added predictive value of classifiers**. Quantify the additional predictive value offered by classifiers compared with standard methods.
5. **Assess classifier performance in targeted settings**. Measures of classifiers' predictive accuracy, such as ROC–AUC, sensitivity and specificity vary depending upon the characteristics of the targeted population (for example, according to prevalence and the proportion of COVID$^+$ and COVID$^-$ individuals that are symptomatic). Where possible, apply the classifier in a test set that reflects the appropriate application setting, for example, by subsampling a test set that is representative of the general population, as we do here.
6. **Examine classifiers' expected utility**. We can specify utilities for each testing outcome, that is, quantify the average benefit accrued from a true positive, the different benefit of a true negative, and similarly the costs attached to false positives and false negatives. Then the expected utility provides a highly context-specific score for quantifying and comparing classifiers' performance.
7. **Out-of-study replication**. Replication studies could be performed in randomly sampled cohorts, or in pilot studies in real-world prediction settings with domain-specific utility functions. There are extra challenges when performing out-of-study replication; in particular, the audio-capture protocols might differ. It would facilitate replication if standardized protocols for audio data gathering are collaboratively developed.

We conclude by outlining some limitations of our study, dataset and findings. There are potentially subtle unmeasured confounders across our recruitment channels REACT and T+T. For example, PCR testing in T+T usually occurs in the days after self-screening of symptoms, whereas PCR tests in REACT are more likely to occur on a date approximately pre-determined by study researchers, and therefore to be independent of participants' symptoms. We attempted to control for such unmeasured confounders by including recruitment channel as one of the matched variables in the test set. Despite matching on measured confounders, some residual predictive variation persists (ROC–AUC = 0.62). Our exploratory approaches for characterizing this residual predictive variation (Methods and Supplementary Note 1) suggest that some of this residual performance may be due to unmeasured confounders, but these results are inconclusive. Our results are suggestive of little utility of ABCS in practice relative to symptoms-based screening (Fig. 5). The development of more sophisticated methods for training audio-based models—such as utilizing voice activity detection, fusing multiple models' predictions, or new augmentation methods—in the presence of symptoms data and recruitment bias is a worthwhile and active area of research which, alongside careful design and replication of studies, will eventually provide full clarity on the potential of ABCS as a tool to protect public health.

Methods

Dataset and study design

This section contains an overview of how the dataset was collected, its characteristics and its underlying study design. More in-depth descriptions are provided in two accompanying papers: Budd and co-workers[23] report a detailed description of the full dataset, whereas Pigoli et al.[30] present the rationale for and full details of the statistical design of our study.

Recruitment sources. Our main sources of recruitment were the REACT study and the NHS T+T system. REACT is a prevalence survey of SARS-CoV-2 that is based on repeated cross-sectional samples from a representative subpopulation defined via (stratified) random sampling from England's NHS patient register[31]. The NHS T+T service was a key part of the UK government's COVID-19 recovery strategy for England. It ensured that anyone developing COVID-19 symptoms could be swab tested, followed by the tracing of recent close contacts of any individuals testing positive for SARS-CoV-2 (ref. 25).

Criteria for enrolment. Enrolment for both the REACT and NHS T+T recruitment channels was performed on an opt-in basis. Individuals participating in the REACT study were presented with the option to volunteer for this study. For the NHS T+T recruitment channel, individuals receiving a PCR test from the NHS T+T pillar 2 scheme were invited to take part in research (pillar 1 tests refer to 'all swab tests performed in Public Health England laboratories and NHS hospitals for those with a clinical need, and health and care workers', whereas pillar 2 comprises 'swab testing for the wider population'[25]). The guidance provided to potential participants was that they should be at least 18 years old, had taken a recent swab test (initially no more than 48 h, changing to 72 h on 14 May 2021), agree to our data privacy statement and have their PCR barcode identifier available, which was then internally validated.

Audio recordings. Participants were directed to the 'Speak up and help beat coronavirus' web page[24]. Here, after agreeing to the privacy statement and completing the survey questions, participants were asked to record four audio clips. The first involved the participant reading out the sentence: 'I love nothing more than an afternoon cream tea', which was designed to contain a range of different vowel and nasal sounds.

This was followed by three successive sharp exhalations, taking the form of a 'ha' sound. The final two recordings involved the participant performing volitional/forced coughs, once, and then three times in succession. Recordings were saved in .wav format. Smart phones, tablets, laptops and desktops were all permitted. The audio recording protocol was homogenized across platforms to reduce the risk of bias due to device types.

Demographic and clinical/health metadata. Existing metadata such as age, gender, ethnicity and location were transferred from linked T+T/REACT records. Participants were not asked to repeat this information to avoid survey fatigue. An additional set of attributes—hypothesized to pose the most utility for evaluating the possibility for COVID-19 detection from audio—was collected in the digital survey. This was in line with General Data Protection Regulation requirements that only the personal data necessary to the task should be collected and processed. This set included the symptoms currently on display (the full set of which are detailed in Fig. 1e,f), and long-term respiratory conditions such as asthma. The participants' first language was also collected to control for different dialects/accents, and complement location and ethnicity. Finally, the test centre at which the PCR was conducted was recorded. This enabled the removal of submissions when cases were linked to faulty test centre results. A full set of the dataset attributes can be found in Budd and colleagues[23].

Final dataset. The final dataset is downstream of a quality control filter (see Fig. 1g), in which a total of 5,157 records were removed, each with one or more of the following characteristics: (1) missing response data (missing a PCR test); (2) missing predictor data (any missing audio files or missing demographic/symptoms metadata); (3) audio submission delays exceeding ten days post test result; (4) self-inconsistent symptoms data; (5) a PCR testing laboratory under investigation for unreliable results; (6) a participant age of under 18; and (7) sensitive personal information detected in the audio signal (see Fig. 3d of ref. 23). Pigoli et al.[30] present these implemented filters in full, and the rationale behind each one. The final collected dataset, after data filtration, comprised 23,514 COVID$^+$ and 44,328 COVID$^-$ individuals recruited between March 2021 and March 2022. Please note that the sample size here differs to that in our accompanying papers, in which Budd et al.[23] reported numbers before the data quality filter was applied, whereas our statistical study design considerations, detailed in a work by Pigoli and colleagues[30], focused on data from the restricted date range spanning March to November 2021. We note the step-like profile of the COVID$^-$ count is due to the six REACT rounds, where a higher proportion of COVID$^-$ participants were recruited than in the T+T channel. As detailed in the geo-plots in Fig. 1a,b, the dataset achieves a good coverage across England, with some areas yielding more recruited individuals than others. We are pleased to see no major correlation between geographical location and COVID-19 status, (Fig. 1c), with Cornwall displaying the highest level of COVID-19 imbalance, with a 0.8% difference in percentage proportion of COVID$^+$ and COVID$^-$ cases.

Data splits. In our pre-specified analysis plan, we defined three training sets and five test sets to define a range of analyses in which we investigate, characterize and control for the effects of enrolment bias in our data:

Randomized train and test sets. A participant-disjoint train and test set was randomly created from the whole dataset, similar to methods in previous works.

Standard train and test set. Designed to be a challenging, out-of-distribution evaluation procedure. Carefully selected attributes such as geographical location, ethnicity and first language are held out for the test set. The standard test set was also engineered to over represent sparse combinations of categories such as older COVID$^+$ participants[30]. The samples included in this split exclusively consist of recordings made prior to 29 November 2021.

Matched train and test sets. The numbers of COVID$^-$ and COVID$^+$ participants are balanced within each of several key strata. Each stratum is defined by a unique combination of measured confounders, including binned age, gender and a number of binary symptoms (for example, cough, sore throat, shortness of breath; see Methods for a full description). The samples included in this split exclusively consist of recordings made prior to 29 November 2021.

Longitudinal test set. To examine how classifiers generalized out-of-sample over time, the longitudinal test set was constructed only from participants joining the study after 29 November 2021.

Matched longitudinal test set. Within the longitudinal test set, the numbers of COVID$^-$ and COVID$^+$ participants are balanced within each of several key strata, similarly as in the matched test set above.

The supports for each of these splits are detailed in Fig. 1h.

Machine learning models

Three separate models were implemented for the task of COVID-19 detection from audio, each representing an independent machine learning pipeline. These three models collectively span the machine learning research space thoroughly—ranging from the established baseline to the current state of the art in audio classification technologies—and are visually represented in Extended Data Fig. 7. We also fitted an RF classifier to predict COVID-19 status from self-reported symptoms and demographic data. The outcome used to train and test each of the prediction models was a participant's SARS-CoV-2 PCR test result. Each model's inputs and predictors, and the details on how they are handled, can be found below. Wherever applicable, we have reported our study's findings in accordance with TRIPOD statement guidelines[32]. The following measures were used to assess model performance: ROC–AUC, area under the precision–recall curve (PR–AUC), and UAR (also known as balanced accuracy). Confidence intervals for ROC–AUC, PR–AUC and UAR are based on the normal approximation method[33], unless otherwise stated to be calculated by the DeLong method[34].

openSMILE–SVM. We defaulted to the widely used openSMILE–SVM approach[35] for our baseline model. Here, 6,373 handcrafted features (the ComParE 2016 set)—including the zero-crossing rate and shimmer, which have been shown to represent human paralinguistics well—are extracted from the raw audio form. These features are then concatenated to form a 6,373-dimensional vector, $f_{\text{openSMILE}}(\mathbf{w}) \rightarrow \mathbf{v}$, where the raw waveform, $\mathbf{w} \in \mathbb{R}^n$ (n = clip duration in seconds × sample rate) is transformed to $\mathbf{v} \in \mathbb{R}^{6,373}$; $\mathbf{v}$ is then normalized prior to training and inference. A linear SVM is fitted to this space and tasked with binary classification. We select the optimal SVM configuration on the basis of the validation set before then retraining on the combined train–validation set.

ResNet-50 BNN. Bayesian neural networks provide estimates of uncertainty, alongside strong supervised classification performance, which is desirable for real-world use cases, especially those involving clinical use. Bayesian neural networks are naturally suited to Bayesian decision theory, which benefits decision-making applications with different costs on error types (for example, assigning unequal weighting to errors in different COVID-19 outcome classifications)[36,37]. We thus supply a ResNet-50 (ref. 38) BNN model. The base ResNet-50 model showed initial strong promise for ABCS[5], further motivating its inclusion in this comparison. We achieve estimates of uncertainty through Monte-Carlo Dropout to achieve approximate Bayesian inference over the posterior, as in ref. 39. We opt to use the pre-trained model for a warm start to the weight approximations, and allow full retraining of layers.

The features used to create an intermediate representation, as input to the convolutional layers, are Mel filterbank features with default configuration from the VGGish GitHub (ref. 40): $\mathbf{X}_i \in \mathbb{R}^{96\times64}$, 64 log-mel spectrogram coefficients using 96 feature frames of 10 ms duration, taken from a resampled signal at 16 kHz. Each input signal was divided into these two-dimensional windows, such that a 2,880 ms clip would produce three training examples with the label assigned to each clip (COVID^{+} or COVID^{-}). Incomplete frames at edges were discarded. As with the openSMILE–SVM, silence was not removed. For evaluation, the mean prediction over feature windows was taken per audio recording, to produce a single decision per participant. To make use of the available uncertainty metrics, Supplementary Note 3 details an uncertainty analysis over all audio modalities for a range of train–test partitions.

SSAST. In recent years, transformers[41] have started to perform well in high-dimensional settings such as audio[42,43]. This is particularly the case when models are first trained in a self-supervised manner on unlabelled audio data. We adopt the SSAST[44], which is on a par with the current state of the art for audio event classification. Raw audio is first resampled to 16 kHz and normalized before being transformed into Mel filter banks. Strided convolutional neural layers are used to project the Mel filter bank to a series of patch level representations. During self-supervised pretraining, random patches are masked before all of the patches are passed to a transformer encoder. The model is trained to jointly reconstruct the masked audio and to classify the order of which the masked audio occurs. The transformer is made up of 12 multihead attention blocks. The model is trained end to end, with gradients being passed all of the way back to the convolutional feature extractors. The model is pre-trained on a combined set of AudioSet-2M (ref. 45) and Librispeech[46], representing over two million audio clips for a total of ten epochs. The model is then fine-tuned in a supervised manner on the task of COVID-19 detection from audio. Silent sections of audio recordings are removed before then being resampled to 16 kHz and normalized. Clips are cut/zero-padded to a fixed length of 5.12 s, which corresponds to approximately the mean length of the audio clip. For cases in which the signal length exceeds 5.12 s (after silence is removed), the first 5.12 s are taken. At the training time, the signal is augmented through applying SpecAugment[47] along with the addition of Gaussian noise. The output representations are mean pooled before being fed through a linear projection head. No layers are frozen and again the model is trained end-to-end. The model is fine-tuned for a total of 20 epochs. The model is evaluated on the validation set at the end of each epoch and its weights are saved. At the end of training the best performing model, over all epochs, is chosen.

Random forest classifier. To predict SARS-CoV-2 infection status from self-reported symptoms and demographic data, we applied an RF classifier with default settings (having self-reported symptoms and demographic data as inputs). In our dataset, predictor variables for the symptoms RF classifier on our dataset comprised: cough; sore throat; asthma; shortness of breath; runny/blocked nose; a new continuous cough; Chronic obstructive pulmonary disease (COPD) or emphysema; another respiratory condition; age; gender; smoker status; and ethnicity. In Han and colleagues' dataset[18], predictor variables for the symptoms RF classifier comprised: tightness of chest; dry cough; wet cough; runny/blocked nose; chills; smell/taste loss; muscle ache; headache; sore throat; short breath; dizziness; fever; runny/blocked nose; age; gender; smoker status; language; and location. Prior to training, categorical attributes were one-hot encoded. No hyperparameter tuning was performed, and models were trained on the combined Standard train and validation sets. For the hybrid symptoms + audio RF classifier, the outputted predicted COVID^{+} probability from an audio-trained SSAST is appended as an additional input variable to the self-reported symptoms and demographic variables listed above.

Matching methodology

The matched test set was constructed by exactly balancing the numbers of individuals with COVID^{+} and COVID^{-} in each stratum where, to be in the same stratum, individuals must be matched on all of (recruitment channel) × (10-year-wide age bins) × (gender) × (all of six binary symptoms covariates). The six binary symptoms matched on in the matched test set were: cough; sore throat; asthma; shortness of breath; runny/blocked nose; and 'at least one symptom'.

Our matching algorithm proceeds as follows. First, each participant is mapped to exactly one stratum. Second, the following matching procedure is applied separately in each stratum: in stratum s (of a total of S strata) let $n_{s,+}$ and $n_{s,-}$ denote the number of individuals with COVID^{+} and COVID^{-}, respectively, and let $\mathcal{A}_{s,+}$ and $\mathcal{A}_{s,-}$ be the corresponding sets of individuals. Use $\mathcal{M}_{s,+}$ and $\mathcal{M}_{s,-}$ to denote random samples without replacement of size $\min\{n_{s,+}, n_{s,-}\}$ from $\mathcal{A}_{s,+}$ and $\mathcal{A}_{s,-}$ respectively. Finally we combine matched individuals across all strata into the matched dataset $\mathcal{M}$ defined as:

$$\mathcal{M} := \cup_{s=1}^{S}(\mathcal{M}_{s,+} \cup \mathcal{M}_{s,-}).$$

The resulting matched test set comprised 907 participants who were COVID positive and 907 who were COVID negative. The matched training set was constructed similarly to the matched test set, though with slightly different strata, so as to increase available sample size. For the matched training set, individuals were matched on all of: (10-year-wide age bins) × (gender) × (all of seven binary covariates). The seven binary covariates used for the matched training set were: cough; sore throat; asthma; shortness of breath; runny/blocked nose; COPD or emphysema; and smoker status. The resulting matched training set comprised 2,599 participants who were COVID positive and 2,599 who were COVID negative.

Quantifying the expected utility of a testing protocol

We consider the action of applying a particular testing protocol to an individual randomly selected from a population. The four possible outcomes $O_{\hat{y},y}$ are

$$O_{\hat{y},y} := [\text{ Predict COVID-19 status as } \hat{y}] \text{ AND } [\text{ True COVID-19 status is } y] \tag{2}$$

for predicted COVID-19 status $\hat{y} \in \{0, 1\}$ and true COVID-19 status $y \in \{0, 1\}$. We denote the probability of outcome $O_{\hat{y},y}$ by

$$p_{\hat{y},y} := \mathbb{P}(O_{\hat{y},y}) \tag{3}$$

and use $u_{\hat{y},y}$ to denote the combined utility of the consequences of outcome $O_{\hat{y},y}$. For a particular population prevalence proportion, π, the $p_{\hat{y},y}$ are subject to the constraints

$$p_{0,1} + p_{1,1} = \pi \tag{4}$$

$$p_{0,0} + p_{1,0} = 1 - \pi, \tag{5}$$

leading to the following relationships, valid for $\pi \in (0, 1)$, involving the sensitivity and specificity of the testing protocol:

$$\text{sensitivity} \equiv \frac{p_{1,1}}{p_{1,1} + p_{0,1}} = \frac{p_{1,1}}{\pi} \tag{6}$$

$$\text{specificity} \equiv \frac{p_{0,0}}{p_{0,0} + p_{1,0}} = \frac{p_{0,0}}{1 - \pi}. \tag{7}$$

The expected utility is:

$$\mathrm{EU} = \sum_{\hat{y}\in\{0,1\}} \sum_{y\in\{0,1\}} u_{\hat{y},y} p_{\hat{y},y} \tag{8}$$

$$= u_{1,1}p_{1,1} + u_{0,1}(\pi - p_{1,1}) + u_{0,0}p_{0,0} + u_{1,0}(1 - \pi - p_{0,0}) \tag{9}$$

$$\begin{aligned} &= \pi[(u_{1,1} - u_{0,1}) \times \text{sensitivity} + u_{0,1}] \\ &+(1 - \pi)[(u_{0,0} - u_{1,0}) \times \text{specificity} + u_{1,0}], \end{aligned} \tag{10}$$

where equations (4) and (5) are substituted into equation (8) to obtain equation (9), and equations (6) and (7) are substituted into equation (9) to obtain equation (10).

Demonstration code

To provide researchers easy access to running the code, we have created a demonstration notebook where the participant is invited to record their own 'sentence', 'cough', 'three cough' or 'exhalation' sounds and evaluate our COVID-19 detection machine learning models on it. The model outputs a COVID-19 prediction, along with some explainable AI analysis, for example, enabling the user to listen back to the parts of the signal which the model allocated the most attention to. In the demonstration, we detail that this is not a clinical diagnostic test for COVID-19, but that it is instead for research purposes and does not provide any medical recommendation, nor should any action be taken following its use. The demonstration file is detailed on the main repository page and can be accessed at https://colab.research.google.com/drive/1Hdy2H6lrfEocUBfz3LoC5EDJrJr2GXpu?usp=sharing.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

To obtain access to 'The UK COVID-19 Vocal Audio Dataset' in full, interested parties may submit their requests to UKHSA at DataAccess@ukhsa.gov.uk. Access is subject to approval and completion of a data sharing contract. For information on how one can apply for UKHSA data, please visit https://www.gov.uk/government/publications/accessing-ukhsa-protected-data/accessing-ukhsa-protected-data. Audio data are provided in .wav format, with four files (one for each recording) for each of the 72,999 participants (unless missing). Metadata are provided in three .csv files, linked by a participant identifier code. Although the dataset is fully anonymized and therefore does not contain any personal data, it has been deposited as safeguarded data in line with the privacy notice provided to participants. Safeguarded data can be used for non-commercial, commercial and teaching projects. To enable wider accessibility, we created another dataset in addition to the original dataset; this dataset has been made open access under a Open Government Licence v3.0. This subset of the 'The UK COVID-19 Vocal Audio Dataset' has been curated to meet the ISB1523: Anonymisation Standard for Publishing Health and Social Care Data standards. Two key changes have been made to achieve this. The 'sentence' modality has been removed; this was non-negotiable from a data privacy perspective as it was classified as personally identifiable information on its own. Furthermore, the granularity of the metadata has been decreased to K3 anonymity after combining all attributes. This was achieved by dropping attributes such as participant location, binning age and date obfuscation. This dataset is available at https://doi.org/10.5281/zenodo.10043978 (ref. 48).

Code availability

The code-base developed for this project can be found at this public GitHub repository: https://github.com/alan-turing-institute/Turing-RSS-Health-Data-Lab-Biomedical-Acoustic-Markers under https://doi.org/10.5281/zenodo.8130844 (ref. 49). Here, instructions are provided to replicate our experimental environment and run our experiments. We have provided a docker image to replicate our experimental set-up fully, which can be initialized with the following command: docker run-it–name-v:/workspace/–gpus=all–ipc=host harrycoppock/ciab:ciab_v4. Further details can be found on the GitHub repository[49].

References

1. Rosengren, H. 14.9 Million excess deaths associated with the COVID-19 pandemic in 2020 and 2021 *World Health Organization* (5 May 2022); https://www.who.int/news/item/05-05-2022-14.9-million-excess-deaths-were-associated-with-the-covid-19-pandemic-in-2020-and-2021
2. Kucharski, A. J. et al. Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study. *Lancet Infect. Dis.* **20**, 1151–1160 (2020).
3. Muller, C. P. Do asymptomatic carriers of SARS-COV-2 transmit the virus? *Lancet Reg.* **4**,100082 (2021).
4. Nessiem, M. A et al. Detecting COVID-19 from breathing and coughing sounds using deep neural networks. In *IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)* https://doi.org/10.1109/CBMS52027.2021.00069 (IEEE, 2021).
5. Laguarta, J., Hueto, F. & Subirana, B. COVID-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open J. Eng. Med. Biol.* **1**, 275–281 (2020).
6. Bagad, P. et al. Cough against COVID: evidence of COVID-19 signature in cough sounds. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2009.08790 (2020).
7. Brown, C. et al. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In *Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 3474–3484 (ACM, 2020); https://doi.org/10.1145/3394486.3412865
8. Imran, A. et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Inform. Med. Unlocked* **20**, 100378 (2020).
9. Pinkas, G. et al. SARS-CoV-2 detection from voice. *IEEE Open J. Eng. Med. Biol.* **1**, 268–274 (2020).
10. Hassan, A., Shahin, I. & Alsabek, M. B. COVID-19 detection system using recurrent neural networks. In *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)* 1–5 (IEEE, 2020).
11. Han, J. et al. Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data. In *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 8328–8332 (IEEE, 2021).
12. Chaudhari, G. et al. Virufy: global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2011.13320 (2021).
13. Lella, K. K. & Pja, A. Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: cough, breath, and voice. *AIMS Public Health* **8**, 240–264 (2021).
14. Andreu-Perez, J. et al. A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test and severity levels. *IEEE Trans. Services Comput.* **15**, 9361107 (2021).
15. Coppock, H. et al. End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study. *BMJ Innov.* **7**, 000668 (2021).
16. Pahar, M., Klopper, M., Warren, R. & Niesler, T. COVID-19 cough classification using machine learning and global smartphone recordings. *Comput. Biol. Med.* **135**, 104572 (2021).

17. Pizzo, D. T. & Esteban, S. IATos: AI-powered pre-screening tool for COVID-19 from cough audio samples. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2104.13247 (2021).
18. Han, J. et al. Sounds of COVID-19: exploring realistic performance of audio-based digital testing. *npj Digit. Med.* **5**, 1–9 (2022).
19. Wynants, L. et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *Br. Med. J.* **369**, m1328 (2020).
20. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
21. Coppock, H., Jones, L., Kiskin, I. & Schuller, B. COVID-19 detection from audio: seven grains of salt. *Lancet Digit. Health* **3**, e537–e538 (2021).
22. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
23. Budd, J. et al. A large-scale and PCR-referenced vocal audio dataset for COVID-19. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2212.07738 (2023).
24. *Speak Up and Help Beat Coronavirus (COVID-19)* (UK Government, 2021); https://www.gov.uk/government/news/speak-up-and-help-beat-coronavirus-covid-19
25. *Department of Health and Social Care (UK), COVID-19 Testing Data: Methodology Note* (UK Government, 2022); https://www.gov.uk/government/publications/coronavirus-covid-19-testing-data-methodology/covid-19-testing-data-methodology-note
26. Murphy, K. P. *Probabilistic Machine Learning: An introduction* (MIT Press, 2022).
27. Stuart, E. A. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* **25**, 1–21 (2010).
28. Kahan, B. C., Forbes, G. & Cro, S. How to design a pre-specified statistical analysis approach to limit p-hacking in clinical trials: the Pre-SPEC framework. *BMC Med.* **18**, 253 (2020).
29. Sah, P. et al. Asymptomatic SARS-CoV-2 infection: a systematic review and meta-analysis. *Proc. Natl Acad. Sci. USA* **118**, e2109229118 (2021).
30. Pigoli, D. et al. Statistical design and analysis for robust machine learning: a case study from COVID-19. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2212.08571 (2022).
31. Chadeau-Hyam, M. et al. REACT-1 study round 14: high and increasing prevalence of SARS-CoV-2 infection among school-aged children during September 2021 and vaccine effectiveness against infection in England. Preprint at *medRxiv* https://www.medrxiv.org/content/early/2021/10/22/2021.10.14.21264965 (2021).
32. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med.* **13**, 1 (2015).
33. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
34. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
35. Eyben, F., Wöllmer, M. & Schuller, B. OpenSmile—the Munich versatile and fast open-source audio feature extractor. In *Proc. 18th ACM International Conference on Multimedia* 1459–1462 (ACM, 2010).
36. Vadera, M. P., Ghosh, S., Ng, K. & Marlin, B. M. Post-hoc loss-calibration for Bayesian neural networks. In *Proc. Thirty-Seventh Conference on Uncertainty in Artificial Intelligence* 1403–1412 (PMLR, 2021).
37. Cobb, A. D., Roberts, S. J. & Gal, Y. Loss-calibrated approximate inference in Bayesian neural networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1805.03901 (2018).
38. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).
39. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proc. 33rd International Conference on Machine Learning*1050–1059 (PMLR, 2016).
40. *Tensorflow/Models* (GitHub, 2019); https://github.com/tensorflow/models/blob/master/research/audioset/vggish/vggish_input.py
41. Vaswani, A. et al. Attention is all you need. In *31st Conference on Neural Information Processing Systems* https://proceedings.neurips.cc/paper_files/paper/2017 (2017).
42. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems* (eds. Lin, H. et al.) Vol. 33, 12449–12460 (Curran Associates, 2020); https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf
43. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*(ICLR, 2021); https://openreview.net/forum?id=YicbFdNTTy
44. Gong, Y., Lai, C.-I. J., Chung, Y.-A. & Glass, J. SSAST: self-supervised audio spectrogram transformer. In *Proc. AAAI Conference on Artificial Intelligence* https://doi.org/10.1609/aaai.v36i10.21315 (AAAI, 2022).
45. Gemmeke, J. F. et al. Audio set: an ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017* (IEEE, 2017).
46. Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 5206–5210 (IEEE, 2015).
47. Park, D. S. et al. SpecAugment: a simple data augmentation method for automatic apeech recognition. In *Proc. Interspeech 2019* 2613–2617 (ISCA, 2019).
48. Coppock, H. et al. *The UK COVID-19 Vocal Audio Dataset (openAccessv1.0)* (Zenodo, 2023); https://doi.org/10.5281/zenodo.10043978
49. Coppock, H. et al. *Alan-Turing-Institute/Turing-RSS-Health-Data-Lab-Biomedical-Acoustic-Markers: Initial* (Zenodo, 2023); https://doi.org/10.5281/zenodo.8130844

Acknowledgements

We gratefully acknowledge the contributions of staff from the NHS Test and Trace Lighthouse Labs, REACT study, Ipsos MORI, Studio24 and Fujitsu Services. Authors at The Alan Turing Institute and Royal Statistical Society Health Data Laboratory gratefully acknowledge funding from the Data, Analytics and Surveillance Group, a part of the UK Health Security Agency (UKHSA). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. We acknowledge The Department for Health and Social Care (grant no. 2020/045) with support from The Alan Turing Institute (grant no. EP/W037211/1) and in-kind support from The Royal Statistical Society; the NIHR Biomedical Research Centre, Oxford (grant no. NIHR203311 to G.N.); the i-sense EPSRC IRC in Agile Early Warning Sensing Systems for Infectious Diseases and Antimicrobial Resistance (grant no. EP/R00529X/1 to J.B.); the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant (grant no. 801604 to A.T.C.); and The Economic and Social Research Council (ESRC) (grant no. ES/P000592/1 to S.P.).

Author contributions

H.C., D.P., K.B., J.B., C.H., S.G.G., V.K., S.R., I.K., G.N., B.W.S. and R.P. conceived and designed the experiments. H.C., J.B., K.B., I.K., G.N., D.H., R.P. and L.B. performed the experiments. H.C., S.P., J.B., K.B., C.H., T.T., J.P., L.B., A.T., S.R., E.K., I.K., G.N., B.W.S., D.H., S.E. and R.P. contributed materials/analysis tools. H.C., D.P., J.B., K.B., I.K., G.N., D.H. and R.P. analysed the data. H.C., K.B., J.B., J.P., V.K., I.K., J.M., S.E., T.T., A.T., S.R., B.W.S., S.G.G., G.N., P.D., R.P., D.P., C.H., L.B., E.K., S.P., D.H. and R.J. wrote the paper.

Competing interests

The authors declare no competing interests.

Ethics statement

This study has been approved by The National Statistician's Data Ethics Advisory Committee (reference NSDEC(21)01) and the Cambridge South NHS Research Ethics Committee (reference 21/EE/0036) and Nottingham NHS Research Ethics Committee (reference 21/EM/0067). All participants reviewed the provided participant information and gave their informed consent to take part in the study.

Additional information

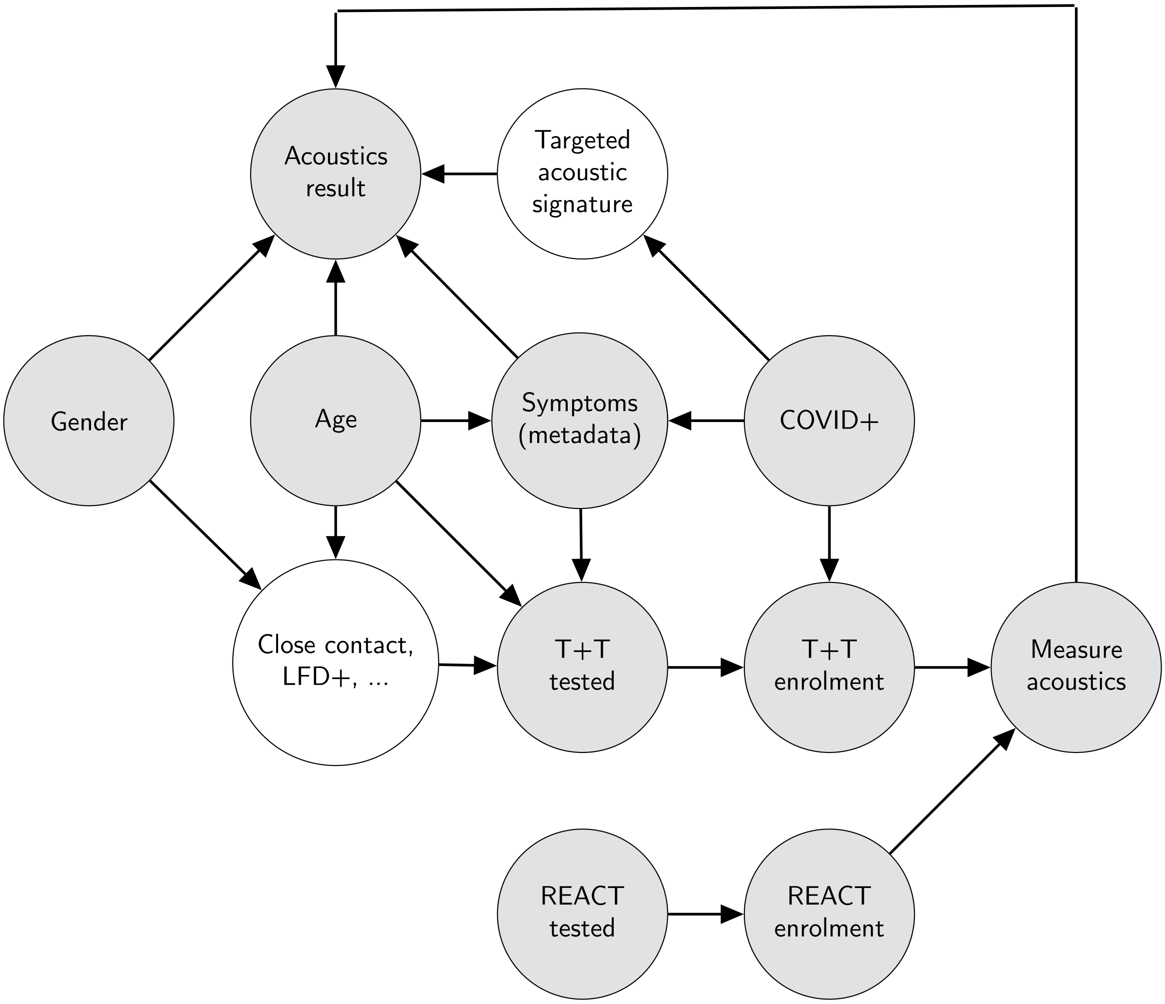
Extended data is available for this paper at https://doi.org/10.1038/s42256-023-00773-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00773-8.

Correspondence and requests for materials should be addressed to Chris Holmes.

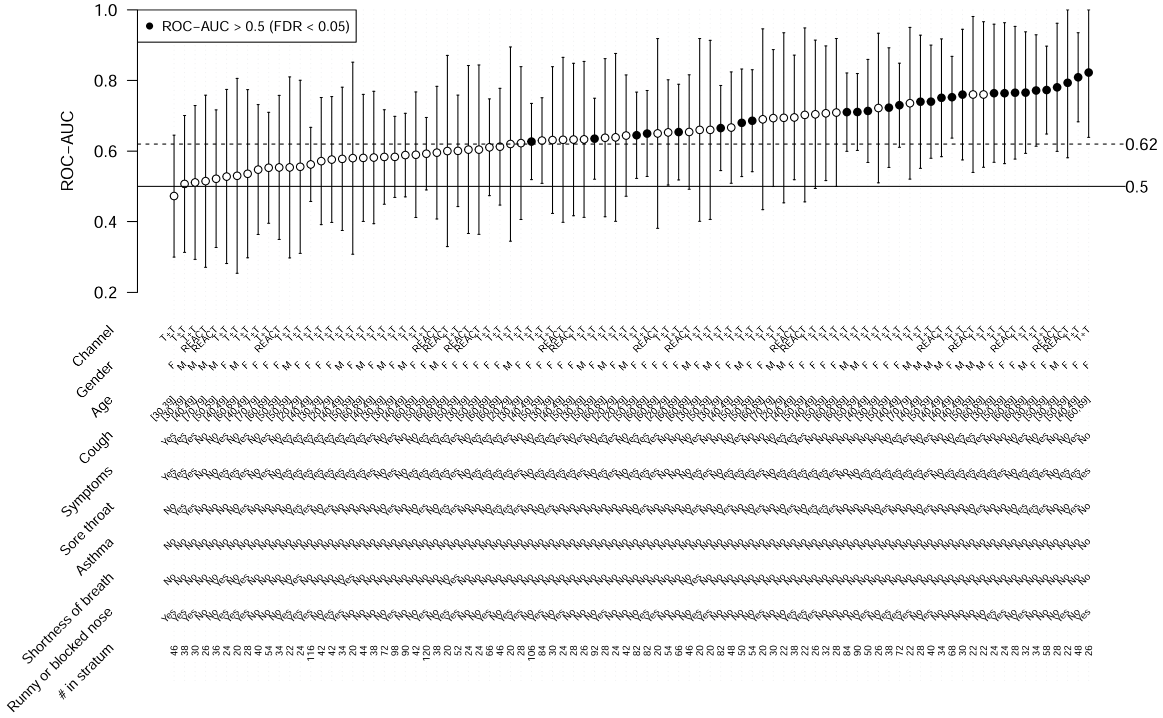
Peer review information *Nature Machine Intelligence* thanks Huijun Ding and Georg Stemmer for their contribution to the peer review of this work.

[1]The Alan Turing Institute, London, UK. [2]Imperial College London, London, UK. [3]University of Oxford, Oxford, UK. [4]NIHR Oxford Biomedical Research Centre, Oxford, UK. [5]Surrey Institute for People-Centred AI, University of Surrey, Guildford, UK. [6]King's College London, London, UK. [7]Division of Medicine, University College London, London, UK. [8]London Centre for Nanotechnology, University College London, London, UK. [9]UK Health Security Agency, London, UK. [10]School of Applied Sciences, University of Brighton, Brighton, UK. [11]Institute of Health Informatics, University College London, London, UK. [12]University of Nottingham, Nottingham, UK. [13]University of Lancaster, Lancaster, UK. [14]University of Augsburg, Augsburg, Germany. [15]These authors contributed equally: Harry Coppock, George Nicholson, Ivan Kiskin. [16]These authors jointly supervised this work: Davide Pigoli, Steven Gilmour, Stephen Roberts, Chris Holmes. ✉e-mail: cholmes@stats.ox.ac.uk

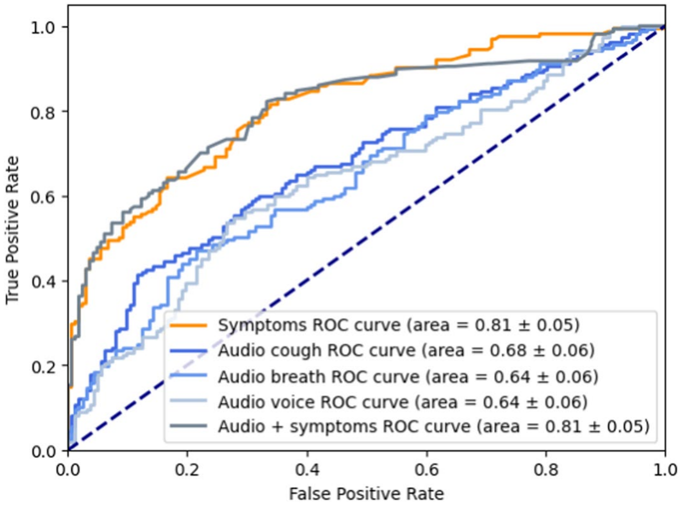


Extended Data Fig. 1 | Bayesian knowledge graph describing the main features of the recruitment process. The nodes in the graph represent the states of an individual in the population; shaded nodes are observed and non-shaded are latent.

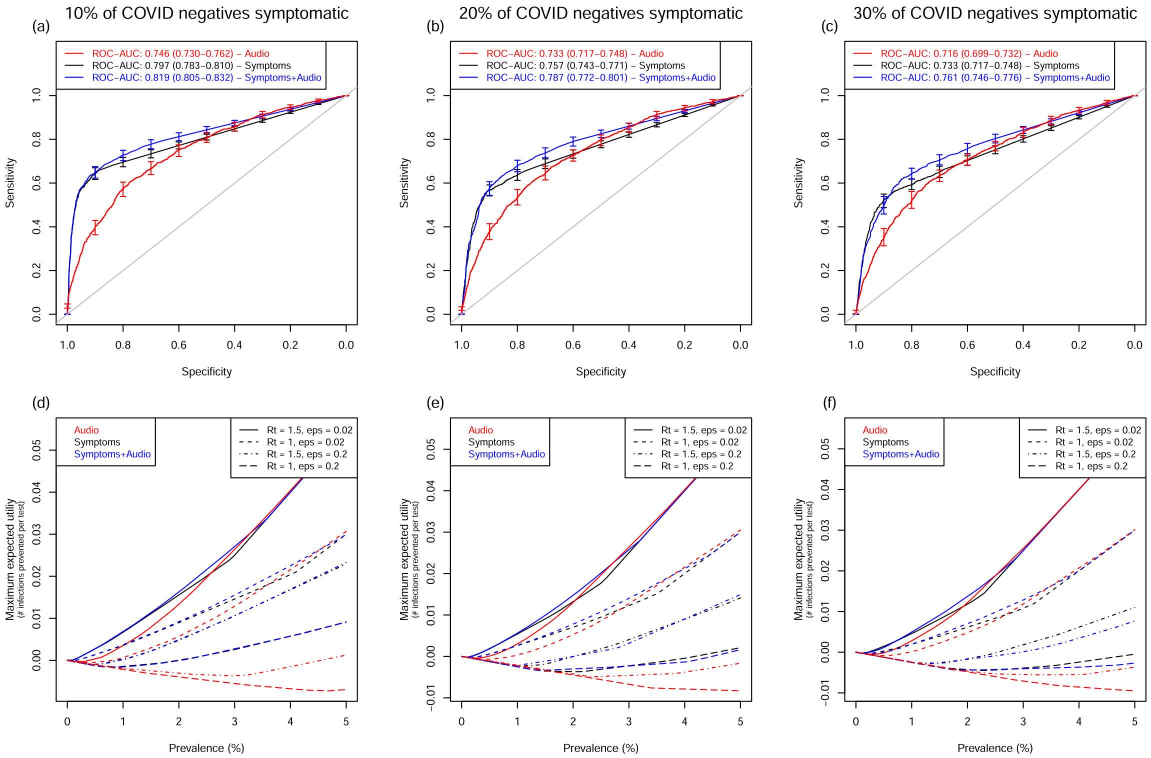
https://doi.org/10.1038/s42256-023-00773-8



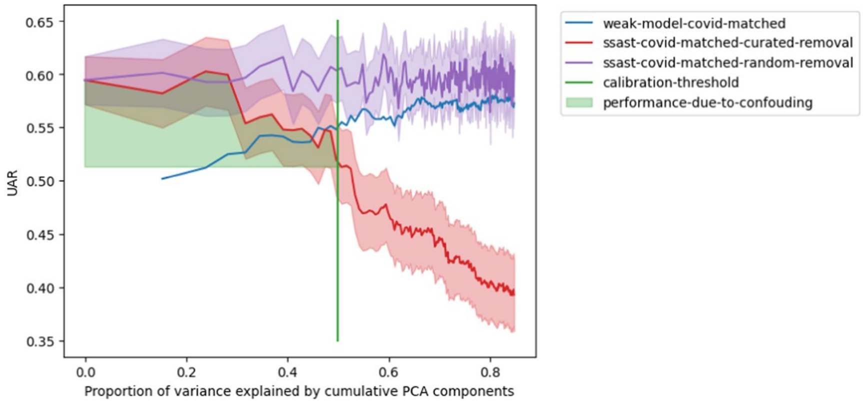
Extended Data Fig. 2 | Predictive accuracy within Matched strata. Estimated ROC–AUC in each of 88 strata in the combined Matched and Longitudinal Matched test sets (these are the 88 largest strata in this combined test set, having at least 10 COVID+ and 10 COVID- participants). Upon controlling False Discovery Rate (FDR) at 5%, we observe significant differences in predictive scores between COVID- and COVID+ individuals in 28 strata (two-tailed Mann-Whitney U test; significance denoted by filled points), suggesting that the classifier has low but consistent predictive power across a large number of strata. Results are presented as estimated ROC–AUC with accompanying error bars denoting DeLong 95% confidence intervals (the sample size underlying each CI is given by '# in stratum' in the final row of the table beneath the plot). Of these CIs, 84 out of 88 (95.4%) are overlapping with 0.62, consistent with a common value of ROC–AUC=0.62 across all strata. Details of each stratum are shown in the table below the plot. The reference value of ROC–AUC=0.62, representing the estimated global (non-stratified) predictive ability of the SSAST classifier (see Table 1) is marked with a horizontal dashed line. The value ROC–AUC=0.5, representative of no predictive ability is marked by a solid horizontal line.



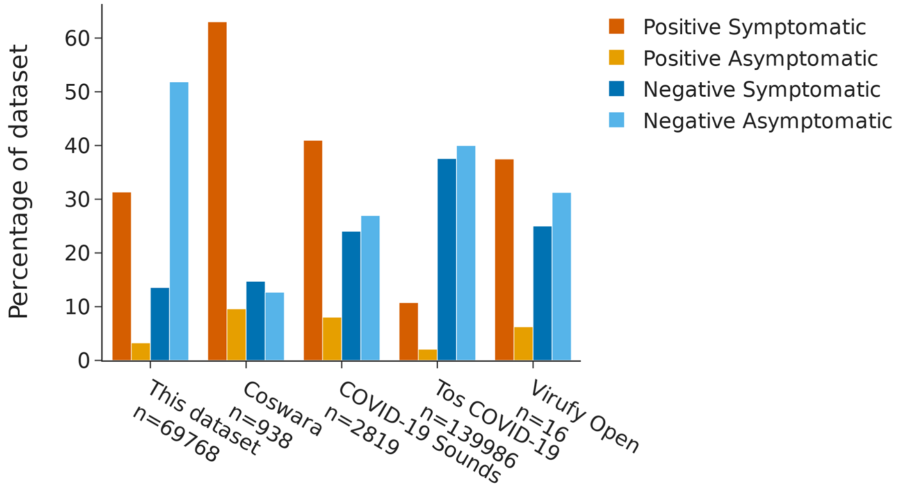
Extended Data Fig. 3 | ROC–AUC SSAST performance when trained and evaluated on the COVID-19 sounds publicly available dataset[18]. Here, our SSAST model's ROC–AUC exceed those of the CNN model of Han et al.[18], but the difference in ROC–AUC between the methods is small and is compatible with random estimation error, as seen from the wide confidence intervals (attributable to the small test set of size 200): cough (ROC–AUC for SSAST 0.68 [0.62-0.74] vs CNN 0.66 [0.60-0.71]), breath (0.64 [0.58-0.70] vs 0.62 [0.56-0.68]), voice (0.64 [0.58-0.70] vs 0.62 [0.56-0.68]). A simple symptoms checker (RF) and a hybrid symptoms-audio are also evaluated for comparison, outperforming both our SSAST audio-only fit and Han et al.'s audio-only CNN.



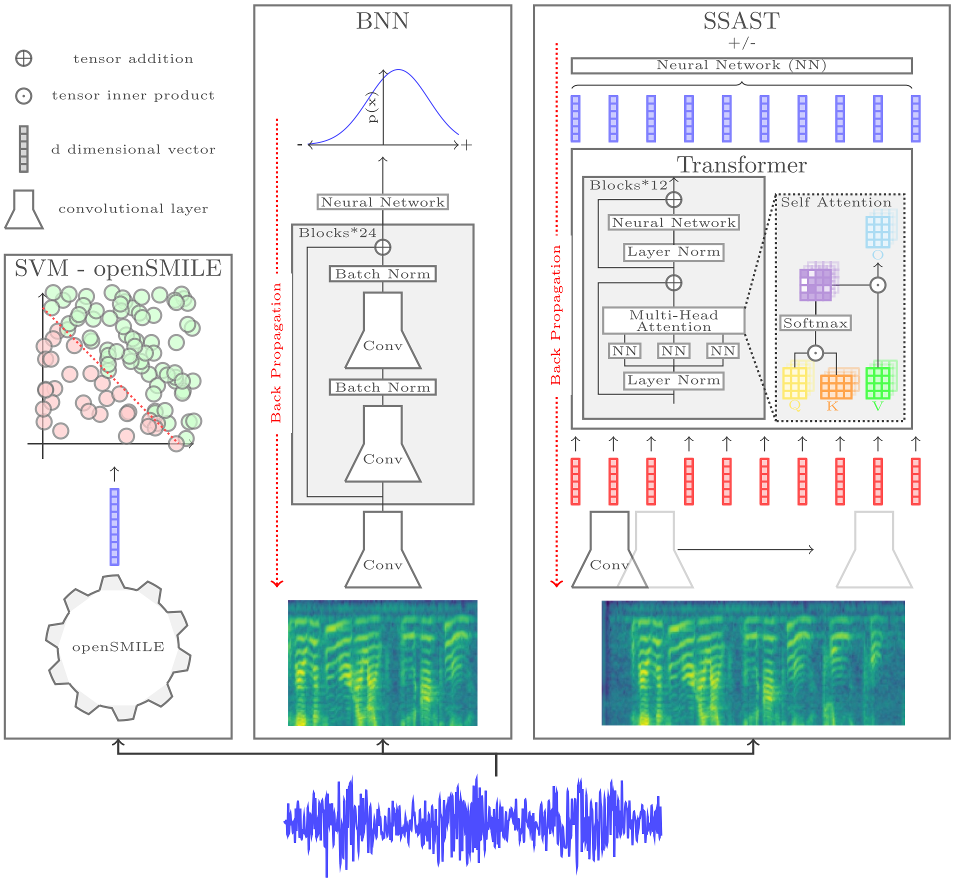
Extended Data Fig. 4 | Comparison of sensitivity, specificity, and utility across audio-based and symptoms-based classifiers, as applied in a simulated general populations. The percentage of COVID- individuals who are symptomatic in the general population varies between 10% and 30% across the three columns of panels (labelled top). (a)-(c) Comparison of ROC curves between the Audio, Symptoms, and Symptoms+Audio classifiers; curves show sensitivity as a function of specificity with error bars denoting 95% CIs for sensitivity, where CIs are calculated in pROC::ci.se and are based on a sample size of 2,000 stratified bootstrap replicates; panel legends show the curve colour for each classifier alongside ROC−AUC estimates and 95% DeLong CIs. (d)-(f) Comparison of maximum expected utility across classifiers. Four different utility functions are included, as detailed in the top-right legend (utility function parameters Rt, ε and δ are defined in Results; in this Figure, $\delta = 0.25$). Curves are coloured to indicate Audio, Symptoms or Symptoms+Audio classifiers, as detailed in the top left legend.



Extended Data Fig. 5 | Results of the Weak-Robust approach. The blue line represents the SVM model trained and evaluated on an increasing number of Principal Component Analysis (PCA) components of openSMILE vector representations of the audio signal for the Matched COVID-19 detection from audio task ('weak-model-covid- matched'). Individuals correctly classified by the weak model in the Matched test set are hypothesized to har- bour confounding signal, and are removed to create the curated Matched test set. The red line shows SSAST performance on this curated Matched test set ('ssast-covid-matched-curated-removal'). For comparison, we also randomly remove Matched test cases and these results are shown by the purple line ('ssast-covid-matched- curated-removal'). The vertical green line corresponds to the calibration threshold, that is, the number of PCs for which the weak model achieves UAR of greater than 80% on the calibration task. The green shaded area corresponds to the drop in SSAST performance that we attribute to the removal of confounding in Matched test set cases. We note that the drop in performance below random classification is hypothesized to be due to only the 'tricky' cases remaining (for example, symptomatic COVID-). The 95% confidence intervals are calculated via the normal approximation method with the outcome of the experiment being the center line.



Extended Data Fig. 6 | Symptomatic vs asymptomatic for other COVID-19 datasets when sufficient information is provided. Coswara[48], COVID-19 Sounds[18], Tos COVID-19[17] and Virufy[49].



Extended Data Fig. 7 | Schematic detailing the three separate pipelines implemented to evaluate ABCS. openSMILE- SVM baseline, the Bayesian Neural Network (BNN) and the Self-Supervised Audio Spectrogram Transformer (SSAST). Both SSAST and BNN first convert the raw audio signal to mel spectrogram space whereas the openSMILE–SVM approach extracts a series of handcrafted audio features.

Extended Data Table 1 | Non-exhaustive record of COVID-19 respiratory audio datasets

Dataset	Positive/Total Participants	Audio Modality	Crowd sourced	COVID-19 Label	Publicly available	Reported Classification metrics		
						F1	ROC-AUC	Acc
Tos COVID-19[16]	25664/139986	Cough	yes	PCR (19.36%), Lateral Flow	yes	0.87	-	-
COVID-19 Sounds[17]	2106†/36116	Cough, Breathing, Speech	yes	Self-reported	yes	-	0.71	
COUGHVID[50]	1010/20072	Cough	yes	Self-reported, Clinician annotation	yes	-	-	0.95‡[51]
Covid19-Cough[52]	682/1324	Cough	yes	PCR (28.85%), Self-reported	yes	-	-	0.96
Coswara[47]	389/2233	Breathing, Cough, Vowel phonation, Speech	yes	Self-Reported	yes	-	0.85‡ [53]	-
Virufy[48]	14/91	Cough	yes	PCR (75.82%), Self-reported	yes	0.93	-	-
NoCoCoDa[54]	10/10	Cough	yes	Self-reported	yes	0.77‡[55]	-	-
CoughDetect[13]	2339†/8380	Cough	no	PCR (100%)	no	-	0.99	-
opensigma[4]	2660/30000	cough	yes	no test required	no	-	-	0.99

† recorded samples NOT number of participants

‡ when the dataset paper does not provide classification metrics then scores are taken from other published work which give classification scores of the corresponding dataset

Publicly available datasets and two highly cited private datasets are shown.

nature portfolio

Corresponding author(s): Chris Holmes

Last updated by author(s): Nov 11, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The online survey "Speak up and help beat coronavirus" was accessible via compatible internet connected devices with the ability to capture audio recordings, such as smartphones, tablets, laptops, and desktop computers.
Data analysis	The accompanying open source repository details the custom code developed to perform the data analysis. No commercial products were used. Python 3.8 served as the main high level language. We have provided a docker image to fully replicate our experimental setup and it can be initialised with the following command: docker run -it --name <name_for_container> -v <location_of_git_repo>:/workspace/ --gpus=all --ipc=host harrycoppock/ciab:ciab_v4 Further details can be found on the GitHub repository which can be found here: https://github.com/alan-turing-institute/Turing-RSS-Health-Data-Lab-Biomedical-Acoustic-Markers

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

To obtain access to the full dataset, named `The UK COVID-19 Vocal Audio Dataset', interested parties may submit their requests to UKHSA at DataAccess@ukhsa.gov.uk. Access is subject to approval and completion of a data sharing contract. For information on how one can apply for UKHSA data, please visit: https://www.gov.uk/government/publications/accessing-ukhsa-protected-data/accessing-ukhsa-protected-data. Audio data are provided in .wav format, with four files (one for each recording) for each of the 72,999 participants (unless missing). Metadata are provided in three .csv files, linked by a participant identifier code.

Although the dataset is fully anonymised, and therefore does not contain any personal data, it has been deposited as safeguarded data in line with the privacy notice provided to participants. Safeguarded data can be used for non-commercial, commercial and teaching projects.

To enable wider accessibility, in addition to the original dataset, we have created another dataset which has been made open access under a Open Government Licence v3.0. This subset of the `The UK COVID-19 Vocal Audio Dataset' has been curated to meet the SB1523: Anonymisation Standard for Publishing Health and Social Care Data standards. To achieve this, two main changes have been made. The `sentence' modality has been removed; this was non-negotiable from a data privacy perspective as it was classified as personally identifiable information on its own. Additionally, the granularity of the metadata has been decreased to K3 anonymity after combining all attributes. This was achieved by dropping attributes such as participant location, binning age, and date obfuscation. This dataset is available at https://zenodo.org/records/10043978, doi:10.5281/zenodo.10043978.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	We use the term "gender" throughout. Gender was determined based on self reporting, and was recognised as a measured confounder due to its potential association with both recruitment and Covid status. To control for recruitment bias, we performed analyses matched on gender and on other measured confounders, such as symptoms and age. Source data, once available, will include disaggregated gender data. In our pre-QC dataset, 59.64% of participants were female (43,537 participants). Participants accepted a privacy statement outlining how their survey and test data would be linked, how their data would be used for research, and made available for re-use by researchers.
Population characteristics	Please see the "Behavioural & social sciences study design" section below.
Recruitment	Participants were recruited in parallel via two SARS-CoV-2 infection testing channels: 1) a community prevalence survey Real-time Assessment of Community Transmission-1 (REACT) and 2) a government testing service NHS Test and Trace (T&T). In each channel, participants were invited to take part in the study after they underwent testing for SARS-CoV-2. Survey responses and audio recordings were linked to their test result. Inclusion criteria in each recruitment channel were: being 18 years of age or older and having a COVID-19 test barcode number. In the REACT channel, participants were randomly selected from National Health Service England records (which include almost the entire population). During the study period, individuals were advised to seek a PCR test through the T&T channel (i) if they were experiencing symptoms; (ii) were identified as a close contact of a Covid case; or (iii) were taking a confirmatory PCR test following a positive lateral flow device (LFD) test. The main recruitment biases affecting our study arise in the T&T channel, primarily from symptoms-based recruitment; also various age/gender combinations may be more or less likely to self-select, and may also have systematically different probabilities of being Covid positive. Our paper focuses on careful treatment of recruitment bias, and uses matching analysis to control for these effects.
Ethics oversight	This study has been approved by The National Statistician's Data Ethics Advisory Committee (reference NSDEC(21)01) and the Cambridge South NHS Research Ethics Committee (reference 21/EE/0036) and Nottingham NHS Research Ethics Committee (reference 21/EM/0067).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Mixed-methods observational study. Participants were recruited on an opt-in basis through two SARS-CoV-2 swab testing programs (REACT and Test-and-Trace) in England between March 2021 and March 2022. Participants were directed to the "Speak up and help beat coronavirus" web page, where they were prompted to record four audio clips (quantitative data), as well as to fill in a symptoms survey (qualitative data). Further existing metadata, such as age (quantitative), and gender, ethnicity, and location (all qualitative), were transferred from linked records in REACT and Test-and-Trace. The study design is described in detail in Methods.
Research sample	The UK Health Security Agency recruited participants voluntarily through the NHS Test and Trace (T&T) programme and the REACT survey in England from March 2021 to March 2022. The median age of participants was 53 years old with 59.64% of participants reporting female gender (43,537 participants), both larger than the UK-wide averages over the corresponding period: UK-wide 2021 median age 40.7 years old, with females comprising 51.0% of the population. The sample is strongly enriched with individuals exhibiting self-identifiable symptoms, as this was one of the criteria for enrolment via the T&T recruitment channel. The rationale for this particular research sample was that participants would have an accompanying PCR test result linked to their self-reported survey data and audio samples.
Sampling strategy	We recruited from REACT and T&T testing channels over a time window, with the rate of accrual of participants determined by self-selection. Recruitment began in March 2021 and ended in March 2022. The eventual overall sample size was 72,999 with matched training and combined test set size of 5,198 and 6,010 respectively. The sampling method used was convenience sampling. Collecting a large participant number was required to ensure that, on matching, sufficient data remained. We adjudged this sample size to be approximately around a minimum of what would be required to train a state-of-the-art audio classification model on a moderately complex task.
Data collection	The online survey 'Speak up and help beat coronavirus' was accessible via compatible internet connected devices with the ability to capture audio recordings, such as smartphones, tablets, laptops, and desktop computers. The participant answered survey questions and provided audio samples in an environment of their choosing, in which third parties and ambient noise could be present. The researchers were not present during data collection. For data analysis the researchers were not blinded to the data (which included participant meta data, SARS-CoV-2 test result, and audio files); nor were researchers blinded to the study hypothesis.
Timing	Data were collected between 2021-03-01 and 2022-03-07.
Data exclusions	A total of 5,157 (out of 72,999) were excluded from the analyses because each of which had one or more of the following characteristics (a) missing response data (missing a PCR test); (b) missing predictor data (any missing audio files or missing demographic/symptoms metadata); (c) audio submission delays exceeding 10 days post test result; (d) self-inconsistent symptoms data; (e) PCR testing lab under investigation for unreliable results; (f) participant age under 18; (g) sensitive personal information detected in audio signal.
Non-participation	Participants were self-selected and so the initialization rate of the survey was 100% participants. Of participants initiating the survey, (68%, 96%) completed the questionnaire and (63%, 85%) completed the questionnaire and the audio recording in (T&T, REACT) respectively. Having completed the questionnaire there was one participant who dropped out and requested that their records be removed.
Randomization	Participants were not allocated into experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging