COVID-19 detection from audio: seven grains of salt

Harry Coppock, Lyn Jones, Ivan Kiskin, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Coppock, Harry, Lyn Jones, Ivan Kiskin, and Björn Schuller. 2021. "COVID-19 detection from audio: seven grains of salt." *The Lancet Digital Health* 3 (9): e537–38. https://doi.org/10.1016/s2589-7500(21)00141-2.

COVID-19 detection from audio: seven grains of salt

Published Online
July 21, 2021
https://doi.org/10.1016/S2589-7500(21)00141-2

Digital mass testing for COVID-19 via a mobile phone application could be made possible through machine learning and its ability to identify patterns in data. COVID-19 appears to confer unique features in the audio produced by infected individuals,[1] and machine learning COVID-19 detection from breath, cough, and speech audio recordings has yielded promising results.[2–4] In this critique, we present seven major issues with this research and argue that further investigation is needed before conclusions about the detectability of COVID-19 from audio can be made. Many of these issues relate to a single question: are the learnt audio representations, which correlate with COVID-19 in the various collected datasets, truly audio biomarkers caused by COVID-19?

One concern is that machine learning algorithms may simply distinguish between healthy individuals and individuals who are generally unwell, rather than detecting COVID-19 itself. Distinguishing between healthy and sick individuals is a much easier but less useful task. Although some researchers have investigated this issue, for example, by constructing tasks in which trained models classify between individuals with other respiratory related diseases and individuals with COVID-19,[2,5] no studies have been able to conclude that COVID-19 itself is truly being identified.

Environmental corruption is another concern. When an audio event, such as a cough, is recorded, information about the surrounding acoustic environment is also included in the audio file. The background noise introduces a potential for bias in the audio dataset. Bias from inadvertent audio additions can exist in many forms; an example in this context could be that COVID-19 positive individuals are more likely to be indoors, perhaps in a medical setting at the time of recording, than those without COVID-19, whereas COVID-19-negative individuals are more likely to be outside or perhaps in a work place setting than those with COVID-19. Each setting has its own unique identifiable environmental audio mode. Any such association can be identified by the machine learning models, as it attempts to make a COVID-19 prediction. This bias can be subtle and, once present in the dataset, is very difficult to detect and remove. This subtlety could explain why it has not been addressed as a limitation in any of the studies so far. Future research should attempt to control the audio environment at the time of recording to address this issue.

Participant awareness, or lack of participant blinding, is an issue. In the majority of the collected datasets, most participants knew their COVID-19 status at the time of recording,[2,3,5–9] which is problematic because information that betrays COVID-19 status can leak into the machine learning models through emotion in the participant's voice and behaviour. As with environmental corruption, participant awareness allows machine learning models to bypass the difficult task of identifying features of true COVID-19 in favour of an easier route to achieving a high classification score; in this case, by associating emotional signals with COVID-19 status.

Another issue is the validity of datasets. Supervised machine learning methods rely on accurate ground truth labelling of the instances. The resultant machine learning tool can only be as accurate as these labels. However, the validity of the labels used in the COVID-19 audio datasets collected so far is questionable because most datasets allow participants to self-report their COVID-19 status and fail to record the type of test participants had;[2,8,9] polymerase chain reaction (PCR) or, the less accurate, lateral flow test. Although some studies do demand a validated PCR test,[6–8,10] the datasets are small and none have been made publicly available at the time of writing. To highlight the severity of this problem, we note that some datasets have accepted self-assessment as a means of labelling the dataset,[3] and others have failed even to detail how COVID-19 status was determined.[5] Additionally, we found little work investigating algorithm sensitivity to PCR test cycle threshold, and thereby viral load.

The availability of codebases and datasets is another cause for concern. An important part of the scientific process is peer review and replication of results. However, in published work on COVID-19 detection from audio, it is rare for teams to make their code base or dataset publicly available, and this has prevented other groups from attempting to replicate their results. Only the authors of four[2,4,8,9] studies have released their code or datasets partially or completely. How can classification metrics on such a sensitive topic be taken seriously if they cannot be cross checked by other research teams?

Comorbidity, geographical, ethnic, and socioeconomic factors are of potential concern in the context of using machine learning to detect COVID-19. The influence of these factors on the spread of COVID-19 is complex. Disease prevalence has been unequal across regions and societal demographics, and patterns in collected datasets are accentuated by an irregular likelihood distribution across groups for participation in these studies. Bias is introduced through resultant opportunities for machine learning models to infer COVID-19 status through association with demographic characteristics.

Lastly, there is a common issue over the level of control that is had over the participant population when developing machine learning models. Artificial intelligence models are proficient in speaker identification. It is therefore imperative that the training, development, and test sets are disjoint participant populations. When participant populations are not controlled, inflated classification scores are reported because the model can easily recognise reappearing participants and classify their COVID-19 status based on cases in the training phase. Nevertheless, several datasets do not record the identity of participants,[9,10] resulting in an inability to avert this issue.

Despite these concerns, evidence suggests that COVID-19 produces identifiable features in infected individuals' speech, cough, and breath audio and progress is being made through the collection of several clinically validated datasets aimed at addressing these seven issues. These datasets will bring us closer to understanding whether the aspiration of an essentially free digital mass test for COVID-19 could become a reality.

HC was responsible for the conceptualisation, literature search, discussion, and writing (of the original draft, its review, and its edit). BS was responsible for supervision, discussion, and writing (of the manuscript's review and its edit). IK and LJ were responsible for the further literature review, discussion, and writing (of the manuscript's review and its edit). We declare no competing interests.

*Harry Coppock, Lyn Jones, Ivan Kiskin, Björn Schuller
harry.coppock@imperial.ac.uk

Department of Computer Science, Imperial College London, London SW7 2RH, UK, (HC, BWS); Radiology Department, North Bristol NHS Trust, Bristol, UK (LJ); Department of Engineering Science, University of Oxford, Oxford, UK (IK)

1 Quatieri T, Talkar T, Palmer J. A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. *IEEE Open Journal of Engineering in Medicine and Biology* 2020; **1:** 203–06.
2 Brown C, Chauhan J, Grammenos A, et al. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. *Proc Data Min Knowl Discov* 2020; 3474–84.
3 Laguarta J, Hueto F, Subirana B. COVID-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open J Eng Med Biol* 2020; **1:** 275–81.
4 Coppock H, Gaskell A, Tzirakis P, Baird A, Jones L, Schuller B. End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study. *BMJ Innov* 2021; **7:** 356–62.
5 Imran A, Posokhova I, Qureshi HN, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Inform Med Unlocked* 2020; **20:** 100378.
6 Pinkas G, Karny Y, Malachi A, Barkai G, Bachar G, Aharonson V. SARS-CoV-2 detection from voice. *IEEE Open J Eng Med Biol* 2020; **1:** 268–74.
7 Pizzo DT, Esteban S, Scetta M. IATos: AI-powered pre-screening tool for COVID-19 from cough audio samples. *arXiv* 2021; published online April 27. https://arxiv.org/abs/2104.13247 (preprint).
8 Bagad P, Dalmia A, Doshi J, et al. Cough against COVID: evidence of COVID-19 signature in cough sounds. *arXiv* 2020; published online Sept 23. https://arxiv.org/abs/2009.08790 (preprint).
9 Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset: a corpus for the study of large-scale cough analysis algorithms. *arXiv* 2020; published online Sept 24. https://doi.org/10.1038/s41597–021–00937–4 (preprint).
10 Andreu-Perez J, Perez-Espinosa H, Timonet E, 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 Serv Comput* 2021; published online Feb 23. https://doi.org/10.1109/TSC.2021.3061402.