

COVID-19 Sound Analysis Project	► Standard Protocol ◀
Section: Data Collection	Version No. 5 Initials: <u>IMM</u>
Title: News & Social Media Videos	Revision Date: 29 June 2022

1. Purpose / Background

1. We hypothesize that subjects who test positive for COVID-19, including asymptomatic subjects, can be accurately discriminated from subjects who test negative for COVID-19 using acoustic biomarkers extracted from speech, coughing, or breathing. If possible, this system can pre-screen for COVID-19 non-invasively, and at essentially zero variable cost.
2. To train a machine learning model, this data collection aspect of the project will focus on gathering recordings of individuals available online—through television, video sharing platforms, social media, or radio—before a positive COVID-19 diagnosis, after a positive COVID-19 diagnosis, and at least 1 month following the positive COVID-19 diagnosis.
3. The Standard Protocol is necessary to ensure that multiple researchers collect audio and video information in a standard and reproducible way.
4. We describe our search strategy from the following dimensions: search scope (time and space), search method, electronic data sources used, search strings, media validation, media documentation, and search results management.

2. Definitions

1. Media
 - i. “Media” in this context refers to any audio content or videos with audio components.
2. Recordings
 - i. “Before recording” in this context is a recording taken at least one month prior to a positive COVID-19 test.
 - ii. “Positive recording” in this context is a recording taken within one week *following* a positive COVID-19 test.
 - iii. “Recovered recording” in this context is a recording taken at least one month following a positive COVID-19 test *and* the subject is not hospitalized *and* not showing symptoms.
3. Clinical
 - i. A “confirmed test” or “positive test” is understood to be a self-reported measure in which the subject confirms having tested positive for COVID-19.
 - ii. “Symptoms” in this context includes any one of the following: cough, difficulty breathing, or runny nose.

3. Procedure

1. Search online through news and social media sites for recordings subjects who have tested positive for coronavirus.

COVID-19 Sound Analysis Project	► Standard Protocol ◀
Section: Data Collection	Version No. 5 Initials: <u>IMM</u>
Title: News & Social Media Videos	Revision Date: 29 June 2022

2. If recording is following a confirmed test, search for recordings of the same subject before the positive test and after recovered (if possible).
3. Only *one* Before, Positive, and Recovered recording *per subject* is to be included.

4. Search Scope

1. All published media related to our search must be within the years 2019 and 2020.
2. There are no geographic limitations to the source of the media related to our search.
3. There are no limitations related to gender.
4. Only subjects above the age of 18 years old will be included.

5. Search Method

1. *Automatic search* involves executing search strings on search engines of electronic data sources.
2. *Manual search* includes media not found through search engine recommendations or with generic search strings. This media may be included so long as it can be easily accessed publicly.
3. The following publishers' sites will be searched:
 - i. YouTube (www.youtube.com)
 - ii. Facebook (www.facebook.com)
 - iii. Instagram (www.instagram.com)
 - iv. Twitter (www.twitter.com)
 - v. TikTok (www.tiktok.com)
 - vi. Fox News (www.foxnews.com)
 - vii. New York Times (www.nytimes.com)
 - viii. NPR (www.npr.org)

6. Search Strings

1. Due to the nature of the media being collected and the attempt to gather multiple recordings from a single subject prior to a positive test, after a positive test, and after recovery, there are no stringent guidelines on search strings.
2. The search strings used in manual searches for subjects must be included in the media documentation file.
3. Suggested search terms

COVID-19 Sound Analysis Project	► Standard Protocol ◀
Section: Data Collection	Version No. 5 Initials: <u>IMM</u>
Title: News & Social Media Videos	Revision Date: 29 June 2022

- i. "COVID-19 interview" + [name]
 - ii. "Covid positive" + [name]
 - iii. "Coronavirus positive interview" + [name]
- 4. Suggested hashtags
 - i. #covidpositive
 - ii. #coronavirus

7. Media Validation

- 1. Inclusion criteria:
 - i. Subject speaks for longer than 4 seconds *consecutively*.
 - ii. If a Before recording, the recording of the subject was taken at least 1 month before Positive recording.
 - iii. If a Positive recording, the recording of the subject was taken within 1 week *following* a positive COVID-19 test.
 - iv. If a Recovered recording, the recording of the subject was taken at least 1 month *following* a positive COVID-19 test *and* the subject is not hospitalized *and* not showing symptoms.
- 2. Exclusion criteria:
 - i. Subject is wearing a mask during recordings.
 - ii. Overwhelming background noise is present in the signal and masks any part of the recording. Background noise *does not* include background music for video segments, so long as the speech, coughing, or breathing are clear and audible.
 - iii. Only media from Before or Recovered is available and no Positive recordings are available.
 - iv. Unable to determine the date of COVID test result.
 - v. Subject is already included (i.e. already accessed recordings from Before, Positive, and Recovered)
 - vi. Link is not permanent (e.g. Instagram stories, Facebook stories, or Fleets which will be unavailable after 24 hours.)

8. Media Documentation

- 1. Search results will be documented on a Google Sheet with the following information:
 - i. Is it a video with audio or audio only?

COVID-19 Sound Analysis Project	► Standard Protocol ◀
Section: Data Collection	Version No. 5 Initials: <u>IMM</u>
Title: News & Social Media Videos	Revision Date: 29 June 2022

- ii. Title of media
 - iii. Subject name (if available in the media)
 - 1. If Subject name is not available, username is acceptable.
 - 2. If neither Subject name nor username are available, an identifier should be generated as such: {First two letters of the website}_{Date posted}_{Date accessed}, for example TW_24Jun20_19Nov20.
 - iv. Is it a Before, Positive, or Recovered recording?
 - v. Which publisher site was the media accessed on?
 - vi. Link to the media
 - vii. Search string used (if used to access media directly)
 - viii. Are there other speakers than the primary subject in the media?
 - ix. If yes, when does the subject *begin* speaking?
 - x. Date media was accessed
2. Following aggregation by researchers, a final selection decision will be made. Decision and exclusion explanation are included in the media documentation.

9. Roles / Responsibilities

- 1. A limited number of researchers will be trained on this procedure. These individuals will be responsible for collecting media and documenting this collection in the appropriate form.
- 2. Ian McLane is responsible for updating the Standard Protocol.
- 3. Drew Grant and Ian McLane are responsible for final inclusion decision.