Protocol X9001293

Acute Respiratory Illness Surveillance (AcRIS) by Monitoring Voice and Illness Symptom Changes using a Mobile application in a Low-Interventional Decentralized Study

Statistical Analysis Plan (SAP)

Version: **5**

Date: **15 July 2022**

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

1. VERSION HISTORY

Table 1. Summary of Changes

Version/ Date	**Associated Protocol Amendment**	**Rationale**	**Specific Changes**
1 25 Mar 2021	Original 19 Jan 2021	N/A	N/A
2 04 Jan 2022	**Amendment #1** *28 September 2021*	Change of Primary Objective and Primary Endpoints	**Change of Primary Objective**: “Obtain data to characterize the relationship between symptoms and voice features for participants with acute viral respiratory illness. This data will be used as the basis to build voice and symptom algorithm(s) for detection and monitoring of these illnesses.” Rationale: Due to current low attack rate of SARS-CoV-2 seen in this study thus far, the primary objective has been modified and expanded to include SARS-CoV-2, influenza virus, and RSV. The primary endpoints now reflect self-reported symptoms and voice changes due to these three viruses. Incorporated changes throughout the SAP body to reflect this protocol update.
		Updated sample size	**Change of Number of Participants**: The total sample size updated from 6250 to approximately 8700 to reflect the updated primary objective of the amendment, the updated overall estimate of the symptomatic attack rate for the 3 viruses, and the updated data attrition rate. Incorporated changes throughout the SAP body to reflect this protocol update.
		Change of Subset Analyses	**Change of Subset Analyses (Section 6.4):** Replaced subset analyses based on Male >60 y/o; Male <=60 y/o; Female >50 y/o; Female <=50 y/o) to subset analyses based on the three viruses (SARS-CoV-2, influenza virus, and RSV).

Table 1. Summary of Changes

Version/ Date	**Associated Protocol Amendment**	**Rationale**	**Specific Changes**
			Added plots of well-to-sick changes by day in symptomatic positive participants for the four gender/age sub-groups.
3 08 Feb 2022	**Amendment #1** *28 September 2021*	Added Appendix A Updated list of acronyms in Appendix B	Appendix A contains the **X9001293 Algorithm Development SAP**, which provides the detailed methodology for summary and statistical analyses related to the development of the screening algorithm for Symptomatic COVID-19 illness using voice features and symptoms data collected in the X9001293 study. Added acronyms related to Algorithm Development SAP.
4 06 May 2022	**Amendment #2** *3 March 2022*	Updated sample size to align with Protocol Amendment #2 Clarified the voice quality descriptions for E-diary voice recordings	**Change of Number of Participants**: CCI [redacted] Incorporated changes throughout the SAP body to reflect this protocol update. **Section 6.2.2.** In order to be consistent with voice quality flags for SNR and signal duration implemented in the study, the SNR and duration thresholds were changed from >= to > such that they SNR > 20 dB for all tasks, and duration > 3 sec for phonemes or > 10 sec for reading). An additional Noise Loud criterion was incorporated to reflect the presence of background noise in the data.

Table 1. Summary of Changes

Version/ Date	Associated Protocol Amendment	Rationale	Specific Changes
			Incorporated changes throughout the SAP body to reflect this clarification, including the language of the secondary endpoint related to voice quality in **Section 2.1, Table 2**.
		Clarified certain secondary analysis sets	The analysis sets for secondary analysis listed in **Section 6.2.1** and **Section 6.2.4.1** were updated to exclude those participants who experienced technical issues due to study operational errors (e.g., received wrong instructions for self-swabs).
		Clarified definition of occurrence of new or increased symptoms	**Section 5.2.2.1.** Provided definition of occurrence of new or increased symptoms, using the participants' initial symptoms in a time window of 3 days.
		Clarified definition of recovery phase	**Section 5.2.2.2.** The SAP defined the beginning of recovery as "the time before 3 consecutive observations of decrease in total symptom scores". This definition was clarified as follows: (a) restrict such decrease to be after the occurrence of new/increased symptoms and after the maximal total symptom score is reached; (b) allow the total symptom score to remain at the same value after the initial drop.
		Added an additional	**Section Interim Analyses Section 7**: This section was updated to include a second interim analysis that may be conducted to

Table 1. Summary of Changes

Version/ Date	Associated Protocol Amendment	Rationale	Specific Changes
		Interim Analysis	evaluate the study primary and secondary objectives, and the algorithm development analyses detailed in appendix SAP based on Cohort 1 data (consisting of approximately 9,100 participants who have been enrolled by 15 Feb 2022 prior to protocol amendment #2).
		Updated intercurrent event wording for estimands	**Section 2.1** and throughout the SAP: the intercurrent event wording for estimands in this study was clarified. There are no particular intercurrent events considered.
		Updated well to sick analysis	**Section 5.2.2.3**: The baseline for well to sick analysis was clarified. To categorize participants during well to sick and facilitate model convergence, the main well to sick analysis was updated to include: 1) a summary was added with respect to those who lack data during the period, non-responders for the endpoint, and responders (for which the days to maximal response and the growth rate will be summarized); 2) a linear mixed effect model was added to evaluate the overall trend for the endpoint; and 3) for endpoints that are statistically significant from 2), the nonlinear sigmoid model was modified to primarily use the logit transformed linear mixed effect model to estimate growth during the period.
		Clarified definition of return to well state	**Section 5.2.2.5**: a clarification for return to well state was added, using the participants' initial 3-day symptoms.

Table 1. Summary of Changes

Version/ Date	Associated Protocol Amendment	Rationale	Specific Changes
		Added a heatmap plot for correlation presentation	**Section 5.2.2.5**: a heatmap plot for the correlation results was added for visualization.
		Added corresponding plot for well to sick descriptive summary	**Section 6.1.1.1**: a corresponding mean/SD plot for the well to sick descriptive summary was added for visualization.
		Added a summary of both symptom and voice compliance for the compliance analysis	**Section 6.2.1**: compliance analysis included a summary for both symptom and voice compliance.
		Added a summary of contents for analysis methods section for continuous endpoints	**Section 5.2.2**: A summary of section contents was added at beginning for clarity.
		Made consistent across sections	**Section 5.2**: Updated the p-value specification for statistical significance in this section from p<0.05 to p<0.10 to be consistent with analyses across sections (such as 90% CI).
		For Appendix A algorithm development SAP:	
		Updated the algorithm	**Appendix A SAP Section 3:** Updated the algorithm analysis sets and well/sick state

Table 1. Summary of Changes

Version/ Date	Associated Protocol Amendment	Rationale	Specific Changes
		analysis sets and well/sick state definitions Added additional algorithm analyses	definitions. Updated the analyses accordingly in **Appendix A SAP Section 5.** **Appendix A SAP Section 4.2.5:** Added that in the classical machine learning modeling a combination of spectrograms and voice features may be explored. **Appendix A SAP Section 4.2.6:** Transformers-based acoustic models may also be explored for deep learning modeling. Added **Appendix A SAP Section 4.3:** on fusion of classical machine learning and deep learning modeling. Added **Appendix A SAP Section 4.4:** an additional well vs. sick analysis with the sick state definition centered around the max total symptom score. Added **Appendix A SAP Section 4.5:** change from baseline analysis. **Appendix A SAP Section 4.6:** Updated on further model testing to incorporate the additional analyses outlined in Section 4. **Appendix A SAP Section 5:** Updated to reflect the additional algorithm analyses.
5 15 July 2022	**Amendment #1** *28 September 2021*	Reverted sample size and descriptions to align with Protocol Amendment #1 CCI [redacted]	**Change of Number of Participants and Protocol descriptions**: The protocol was reverted from Amendment #2 back to Amendment #1; as such the total sample size was reverted to approximately 8,700, and changes were incorporated throughout the SAP body to reflect this protocol reversion. CCI [redacted]

Table 1. Summary of Changes

Version/ Date	Associated Protocol Amendment	Rationale	Specific Changes
			CCI
		For Appendix A algorithm development SAP:	
		Updated the analysis sets. (i) Removed Primary Analysis Set 1 - since protocol amendment #2 was not carried out, there was insufficient matching data to generate this analysis set and conduct the corresponding analysis. (ii) Included influenza virus, and RSV in Secondary Analysis Set B	**Appendix A SAP Section 3:** Updated the analysis sets. Updated the analyses accordingly in **Appendix A SAP Section 5.**

Table 1. Summary of Changes

Version/ Date	Associated Protocol Amendment	Rationale	Specific Changes
		and CCI to incorporate all RT-PCR positive cases to match the primary analysis.	
		Updated machine learning and deep learning output performance measures.	**Section 4.2.5.2., Section 4.2.6.4.** CCI Performance measures of interest: *sensitivity*, *specificity* and *balanced accuracy*. Dropped *F1-score, positive predicted value and negative predicted value* since these are sample-based measures that depend on arbitrary enrollment ratio.
		Updated the classical machine learning approach and the change from baseline approach to include an additional analysis method - the covariance approach.	**Section 4.2.5. and 4.5.** Added description of the Covariance Approach.
		CCI	CCI

Table 1. Summary of Changes

Version/ Date	Associated Protocol Amendment	Rationale	Specific Changes
		CCI [redacted]	

2. INTRODUCTION

Text taken directly from the protocol is *italicized.*

The purpose of the AcRIS study is to obtain data to characterize the relationship between symptoms and voice features for RT-PCR confirmed SARS-CoV-2, influenza virus, or RSV positive participants with acute viral respiratory illness. This data will be used as the basis to build voice and symptom algorithm(s) for detection and monitoring of these illnesses. This would benefit vaccine development across several key disease areas, including RSV and influenza, beyond the immediate application to SARS-CoV-2.

The study also models concepts of more efficient "flexible" trials involving not only voice recognition, but also web-based participant recruitment, enhanced participant engagement, and remote sample collection that could make future clinical studies more efficient. The clinical data obtained in this observational study could provide the documentation of the technology's performance needed to enable its deployment in future interventional studies.

Overall Design: This is a low-interventional, observational study with planned enrollment of approximately 8,700 participants 18 years of age or older. Participants will record acute respiratory illnesses and SARS-CoV-2 symptoms and voice data daily for up to a maximum of 8 weeks in both the well state and, should they become ill, the sick state, utilizing the Electronic diary on their Mobile application. Once enrolled, the participant will start recording symptoms and voice in the Electronic diary, with daily time commitment to this portion of the study expected to be 2-4 minutes. Two nasal self-swab collection kits will be ordered for delivery to the participant once they are enrolled in the study. The participant will be asked to self-swab when the test kit arrives (swab #1). The kit, including the specimen, will be returned to the central lab for SARS-CoV2/Influenza/RSV RT-PCR testing. The participant is expected to complete 3 phonemes and 5 lines of reading each day, in addition to score the self-reported symptoms in the Electronic diary. If participants become sick (self-report) with new or increased symptoms of respiratory illness, they will be asked to self-swab (swab #2) and return the sample for central SARS-CoV-2/Influenza/RSV RT-PCR *testing. If the participant does not develop any new or increased symptoms between swab #1 and end of Week 6, they will obtain a self-swab (swab #2) at Day 42.*

If the participant tests positive for any of the three viruses at swab #1 or swab #2, they will continue the study until end of week 8. If they test negative for any of the three viruses at swab #1 and swab #2, they will exit the study at approximately the end of week 6. Results of the RT-PCR testing will be shared with participants.

Demographic, medical history, and smoking status data will be collected. At the end of the study, voice and symptom trajectories will be analyzed to understand the SARS-CoV-2 and other infectious respiratory disease trajectory. Analysis of symptom and voice changes will be examined based on SARS-CoV-2/Influenza/RSV test results. The study design purely contemplates capture of symptoms and voicing by a research participant. It does not involve in any respect the assessment, examination, diagnosis, prognostication or treatment of any study participant.

This statistical analysis plan (SAP) provides the detailed methodology for summary and statistical analyses of the data collected in Study X9001293. This document may modify the plans outlined in the protocol; however, any major modifications of the primary endpoint definition or its analysis will also be reflected in a protocol amendment.

2.1. Study Objectives, Endpoints and Estimands

Table 2. Study Objectives, Endpoints and Estimands

Objectives	***Endpoints***	**Estimands**
Primary Objective(s):	***Primary Endpoint(s):***	**Primary Estimand:**
• *Obtain data to characterize the relationship between symptoms and voice features for participants with acute viral respiratory illness. This data will be used as the basis to build voice and symptom algorithm(s) for detection and monitoring of these illnesses.*	• *Change in self-reported symptom scores in the Electronic diary from well to sick in symptomatic participants with RT-PCR confirmed SARS-CoV-2, influenza virus, or RSV.* • *Change in voice features, such as pitch, jitter, harmonicity, entropy, flatness, shimmer from the voice collection as captured by the Electronic diary from well to sick in symptomatic participants with RT-PCR confirmed*	For this primary objective, analyses will include modeling of well to sick changes for symptoms and voice features in symptomatic SARS-CoV-2, Influenza virus or RSV positive participants, and evaluating the well state baseline across all participants who are initially SARS-CoV-2/ Influenza /RSV negative and have no acute symptoms. Estimand E1: This estimand is intended to provide a population level estimate of the well state as well as

Table 2. Study Objectives, Endpoints and Estimands

Objectives	***Endpoints***	**Estimands**
	SARS-CoV-2, influenza virus, or RSV.	changes in symptoms and voice features from well to sick in symptomatic SARS-CoV-2, Influenza virus or RSV positive participants using Electronic diary. Population: Participants who are ≥18 years of age as defined by the inclusion and exclusion criteria, initially are SARS-CoV-2/ Influenza /RSV negative and have no acute symptoms, and are at risk of developing symptomatic SARS-CoV-2, Influenza virus or RSV illness. Intercurrent events: • There are no particular intercurrent events considered. Participants who have events that lead to missing data will be included and missing data will not be imputed. • Inadequate compliance – data will be used as recorded.
Secondary Objective(s):	***Secondary Endpoint(s):***	**Secondary Estimand:**
1. *Assess compliance of the participants using the Electronic diary to collect data.*	• *Percentage of total days of symptoms entered in the Electronic diary.* • *Percentage of total days of voice recordings entered in the Electronic diary.*	Estimand E2: This estimand is intended to provide a population level estimate of the percentage of occurrence for these endpoints. Population: Participants who are ≥18 years of age as

Table 2. Study Objectives, Endpoints and Estimands

Objectives	***Endpoints***	**Estimands**
2. *Test the quality of the recording from the Electronic diary to collect data that is usable for interpretation.*	• Percentage of quality voice recordings based on signal to noise ratio, sound duration and background noise measures.	defined by the inclusion and exclusion criteria. Intercurrent events: • There are no particular intercurrent events considered. Participants who have events that lead to missing data will be included and missing data will not be imputed. • Inadequate compliance – data will be used as recorded.
3. *Determine rates of positivity for SARS-CoV-2/Influenza/RSV infection.*	• *Percentage of participants with SARS-CoV-2; and/or Influenza; and/or RSV RT-PCR based positivity in self-swabs.*	
4. *Evaluate the feasibility of obtaining self-swabs.*	• *Percentage of participants administering the self-swab at self-swab #1 and self-swab #2.* • *Percentage of participants reporting symptoms in the Electronic diary, who have a self-swab collected at or around symptom onset.* • *Percentage of self-swabs with valid (positive or negative) or invalid (non- reportable due to technical or self-collection failures) results.*	
CCI [redacted]	[redacted]	
[redacted]	[redacted]	

Table 2. Study Objectives, Endpoints and Estimands

Objectives	***Endpoints***	**Estimands**
CCI [redacted]	[redacted]	CCI [redacted]
CCI [redacted]	[redacted]	
[redacted]	[redacted]	[redacted]

Table 2. Study Objectives, Endpoints and Estimands

Objectives	***Endpoints***	**Estimands**
	CCI [redacted]	[redacted]
Safety Objective(s):	***Safety Endpoint(s):***	
To describe the safety profile of self-administered nasal swabs for SARS-CoV-2/Influenza/RSV RT-PCR among participants.	*Incidence and severity of adverse events (AEs)/research related injuries (RRIs) among all enrolled participants.*	No defined estimand for the safety endpoints. These endpoints will be analyzed using Pfizer data standards as applicable.

2.1.1. Primary Estimand(s)

The primary estimand is intended to provide a population level estimate of the well state as well as changes in symptoms and voice features from well to sick in symptomatic SARS-CoV-2, Influenza virus or RSV positive participants using Electronic diary. This estimand (Estimand E1) is defined according to the primary objective and is in alignment with the primary endpoints. It includes the following 4 attributes:

- Population: Participants who are ≥18 years of age as defined by the inclusion and exclusion criteria, initially are SARS-CoV-2/Influenza/RSV negative and have no acute symptoms, and have potential to develop symptomatic SARS-CoV-2, Influenza virus or RSV illness.

- Variables: (1) Change in self-reported symptom scores in the Electronic diary from well to sick in symptomatic SARS-CoV-2, Influenza virus or RSV positive participants; (2) Change in voice features, such as pitch, jitter, harmonicity, entropy, flatness, shimmer from the voice collection as captured by the Electronic diary from well to sick in symptomatic SARS-CoV-2, Influenza virus or RSV positive participants.

- Intercurrent event(s):
  - There are no particular intercurrent events considered. Participants who have events that lead to missing data will be included and missing data will not be imputed.
  - Inadequate compliance – participants' data will be used as recorded.

- Population-level summary: The well state characteristics as well as changes in symptoms and voice features from well to sick in symptomatic SARS-CoV-2, Influenza virus or RSV positive participants using Electronic diary.

2.1.2. Secondary Estimand(s)

The secondary estimand of this study is to provide a population level estimate of the percentage of occurrence for the compliance, quality of voice recordings, infection rates and feasibility of self-swabbing. This estimand (Estimand E2) is defined according to the secondary objectives and is in alignment with the secondary endpoints. It includes the following 4 attributes:

- Population: Participants who are ≥18 years of age as defined by the inclusion and exclusion criteria.

- Variables: All secondary endpoints, when appropriate.

- Intercurrent event(s):
  - There are no particular intercurrent events considered. Participants who have events that lead to missing data will be included and missing data will not be imputed.
  - Inadequate compliance – participants' data will be used as recorded.

- Population-level summary: The percentage of occurrence for the endpoints.

CCI

[redacted]

- CCI [redacted]

[redacted]

[redacted]

[redacted]

2.2. Study Design

This is a low-interventional, observational study with planned enrollment of approximately 8,700 participants 18 years of age or older in order to collect a total N=100 participants with (1) RT-PCR confirmed negative SARS-CoV-2, influenza virus and RSV (swab #1) and (2) RT-PCR confirmed positive SARS-CoV-2, influenza virus and RSV (swab #2 or any subsequent swab) symptomatic completers.

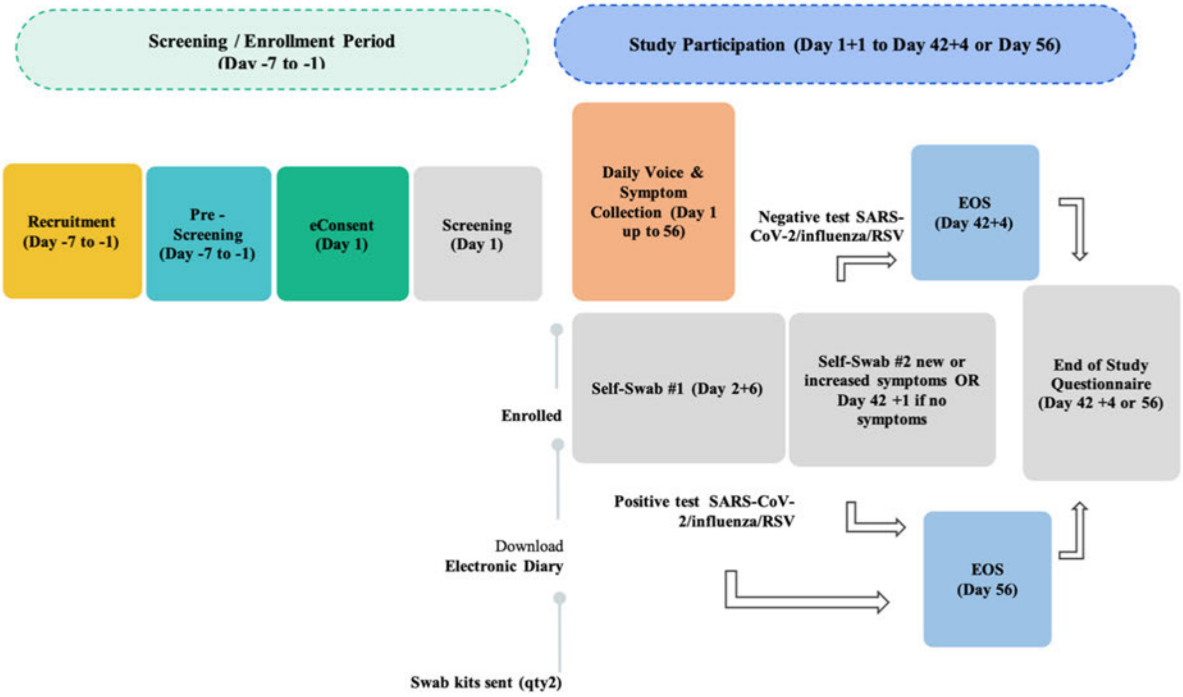
Each participant will be required to stay in the study for 6 weeks. If the participant tests positive for any of the three viruses at swab #1 or swab #2, they will continue the study until the end of Week 8.

Participants will record acute respiratory illnesses and SARS-CoV-2 symptoms and voice data daily for up to a maximum of 8 weeks in both the well state and, should they become ill, the sick state, utilizing the Electronic diary on their Mobile application. Once enrolled, the participant will start recording symptoms and voice in the Electronic diary, with daily time commitment to this portion of the study expected to be 2-4 minutes. Two nasal self-swab collection kits will be ordered for delivery to the participant once they are enrolled in the study. The participant will be asked to self-swab when the test kit arrives (swab #1). The kit, including the specimen, will be returned to the central lab for RT-PCR SARS-CoV2/Influenza/RSV RT-PCR testing. The participant is expected to complete 3 phonemes and 5 lines of reading each day, in addition to score the self-reported symptoms in the Electronic diary. If participants become sick (self-report) with new or increased symptoms of respiratory illness, they will be asked to self-swab (swab #2) and return the sample for central SARS-CoV-2/Influenza/RSV RT-PCR testing. If the participant does not develop any *new or increased symptoms between swab #1 and end of Week 6, they will obtain a self-swab (swab #2) at Day 42.*

If the participant tests positive for any of the three viruses at swab #1 or swab #2, they will continue the study until end of week 8. If they test negative for the three viruses at swab #1 and swab #2, they will exit the study at approximately the end of week 6 when the test results are returned. The results of the RT-PCR testing will be shared with participants.

Demographic, medical history, and smoking status data will be collected. At the end of the study, voice and symptom trajectories will be analyzed to understand the SARS-CoV-2/influenza virus/RSV respiratory disease trajectories. Analysis of changes in symptoms and voice characteristics will be stratified by SARS-CoV-2/Influenza/RSV test results.

Figure 1. Schema for AcRIS Study Design


Screening / Enrollment Period
(Day -7 to -1)
Study Participation (Day 1+1 to Day 42+4 or Day 56)
Recruitment (Day -7 to -1)
Pre - Screening (Day -7 to -1)
eConsent (Day 1)
Screening (Day 1)
Daily Voice & Symptom Collection (Day 1 up to 56)
Negative test SARS-CoV-2/influenza/RSV
EOS (Day 42+4)
Enrolled
Self-Swab #1 (Day 2+6)
Self-Swab #2 new or increased symptoms OR Day 42 +1 if no symptoms
End of Study Questionnaire (Day 42 +4 or 56)
Positive test SARS-CoV-2/influenza/RSV
Download Electronic Diary
EOS (Day 56)
Swab kits sent (qty2)

3. ENDPOINTS AND BASELINE VARIABLES: DEFINITIONS AND CONVENTIONS

3.1. Primary Endpoint(s)

The primary endpoints for this study are:

(a) Change in self-reported symptom scores in the Electronic diary from well to sick in symptomatic participants with RT-PCR confirmed SARS-CoV-2, influenza virus, or RSV.

(b) Change in voice features, such as pitch, jitter, harmonicity, entropy, flatness, shimmer from the voice collection as captured by the Electronic diary from well to sick in symptomatic participants with RT-PCR confirmed SARS-CoV-2, influenza virus, or RSV.

3.2. Secondary Endpoint(s)

The endpoints linked to each secondary objective are as follows:

Endpoints for secondary objective 1: *(a) Percentage of total days of symptoms; (b) Percentage of total days of voice recordings entered in the Electronic diary.*

Endpoints for secondary objective 2: Percentage of quality voice recordings based on signal to noise ratio, sound duration and background noise measures.

Endpoints for secondary objective 3: *Percentage of participants with SARS-CoV-2; and/or Influenza; and/or RSV RT-PCR based positivity in self-swabs.*

Endpoints for secondary objective 4: *(a) Percentage of participants administering the self-swab at self-swab #1 and self-swab #2. (b) Percentage of participants reporting symptoms in the Electronic diary, who have a self-swab collected at or around symptom onset. (c) Percentage of self-swabs with valid (positive or negative) or invalid (non-reportable due to technical or self-collection failure) results.*

CCI

[redacted]

[redacted]

[redacted]

3.4. Baseline Variables

Demographic variables will be summarized. Age and gender may be used to investigate symptoms and voice features in four groups (Male >60 y/o; Male <=60 y/o; Female >50 y/o; Female <=50 y/o). Race may be included as a covariate in the analyses.

The enrollment questionnaire will be summarized. The following components of the enrollment questionnaire may be considered to stratify groups or be included as covariates:

- Current symptoms associated with respiratory infection, such as cold and flu; asthma and allergies

- Pre-existing medical conditions (yes/no)
- Current symptoms associated with pre-existing conditions (yes/no)
- Smoking status and frequency

3.5. Safety Endpoints

Safety profile of self-administered nasal swabs for SARS-CoV2/Influenza/RSV RT-PCR among participants: Incidence and severity of adverse events (AEs)/research related injuries (RRIs) among all enrolled participants.

4. ANALYSIS SETS (POPULATIONS FOR ANALYSIS)

Data for all participants will be assessed to determine if participants meet the criteria for inclusion in each analysis population prior to releasing the database and classifications will be documented per standard operating procedures.

Population	Description
Full Analysis Set (FAS)	All participants who sign the informed consent document and enroll in the study. *Qualified participants are males or females aged 18 years and older (or the minimum state-specific age of consent if > 18) at Screening visit, who are willing and able to comply with daily symptom and voice assessments on the Electronic diary application and other study procedures including self-collection of nasal swabs, are expected to be available for the duration of the study, and are capable of giving informed consent.*
Primary Analysis set	Participants in FAS who initially are SARS-CoV-2, Influenza and/or RSV negative and have no acute symptoms.
CCI	[redacted]

5. GENERAL METHODOLOGY AND CONVENTIONS

An interim analysis will be performed once the first 1000 participants have completed the study (See Section 7). The study will continue whilst the interim analysis is being conducted.

A second interim analysis will be performed on Cohort 1 data, consisting of approximately 9100 participants enrolled by 15 Feb 2022 prior to Protocol Amendment 2. The study will continue whilst the second interim analysis is being conducted.

The primary analysis will be performed after the study is completed and final database is locked.

5.1. Hypotheses and Decision Rules

This is an exploratory study primarily intended to collect data to characterize the relationship between symptoms and voice features for participants with acute viral *respiratory illness. The data will be used to quantify changes from well to sick using voice and self-reported symptoms collected with an Electronic diary in symptomatic SARS-CoV-2/influenza virus/RSV positive participants. The hypotheses that voice and self-reported symptoms will change with acute viral respiratory illness will be tested. This data will be used as the basis to build voice and symptom algorithm(s) for detection and monitoring of these illnesses.* No decision rules will be made.

5.2. General Methods

Data will be summarized using descriptive statistics (number of participants (n), mean, median, standard deviation (SD), minimum and maximum, as appropriate) for continuous (or near continuous) variables, and using frequency and percentages for discrete variables.

Participants can complete the Electronic diary multiple times per day, and each time is considered a session. The session with the most compliant voice and symptom data, and the highest quality voice data will be used for that day to compute the summary statistics. All individual sessions will be used for statistical modeling.

P-values will be generated where appropriate and any p-value < 0.10 will be considered statistically significant. Due to the exploratory nature of the study, p-values will not be adjusted for multiplicity. P-values will be rounded to 3 decimal places and therefore presented as 0.xxx; P-values smaller than 0.001 will be reported as ‘<0.001’.

5.2.1. Analyses for Binary Endpoints

For binary endpoints we will derive summary percentages and the corresponding descriptive statistics, as appropriate.

5.2.2. Analyses for Continuous Endpoints

Log transformation may be performed for continuous endpoints as appropriate. Voice task data that fail in technical QC will be excluded in analysis as appropriate.

This section provides analysis methods for continuous endpoints in various well to sick, well state, [redacted] analyses. It is organized as follows:

- 5.2.2.1 defines occurrence of new or increased symptoms (symptom onset)
- 5.2.2.2 defines occurrence of recovery phase
- 5.2.2.3 provides methods for well to sick analysis, which covers participant’s sickness progression period from symptom onset to before recovery
- 5.2.2.4 provides methods for well state analysis, which covers participant’s well state period from study start to before symptom onset (or to study end if no symptom onset is found)
- 5.2.2.5 provides methods for correlation analysis between symptoms and voice features, which uses participant’s data from symptom onset to the time before return to well state (the latter of which will be defined in this section)

- CCI [redacted]

5.2.2.1. Definition of Occurrence of New or Increased Symptoms

To define the first occurrence of new or increased symptoms for each participant who has no acute symptoms at enrollment, firstly the maximal total symptom score during the first 3 days with available symptom data of the participant is identified (defined as the Initial Symptom Score). The total symptom score is the sum of all the symptom responses per session converted to numeric as described in Section 5.2.3. An observation is defined as the mean of the daily total symptom score across sessions for each compliant day.

Then the first occurrence of new or increased symptoms is defined as the first day after these 3 days when the observed total symptom score is greater than the Initial Symptom Score.

If no greater value in total symptom score is observed after the first 3 days with available symptom data, no occurrence of new or increased symptoms is defined, and the participant is considered as asymptomatic.

5.2.2.2. Definition of Recovery Phase

To define the recovery phase for each participant who has new or increased symptoms (as defined in Section 5.2.2.1), firstly the maximal total symptom score on or after the first occurrence of new or increased symptoms, as well as the first day on which this maximal total symptom score is reached (defined as the Maximal Score Day), are identified.

Then the beginning of recovery is defined as the first day after the Maximal Score Day when the total symptom score is decreased from the previous day with available symptom data, followed by two more days of non-increasing total symptom scores.

When no beginning of recovery is observed, all days up to the final study day of the participant are considered as before recovery.

5.2.2.3. Well to Sick Analysis

For continuous endpoints during well to sick, i.e. after first occurrence of new or increased symptoms in participants based on Electronic diary (defined in Section 5.2.2.1), the change from baseline will be summarized descriptively by day, defining the first occurrence of new or increased symptoms based on Electronic diary as day=1. For the purpose of the well to sick analysis, the baseline is defined as the average of the endpoint values during the 7-day period before the first occurrence of new or increased symptoms. If a participant has no data during these 7 days, the average of the endpoint values for the closest 3 available days prior to these 7 days will be used for the baseline.

The disease progression from well to sick will be modelled monotonically using data only up to the time before recovery, so that the potential recovery phase afterwards is excluded. The time before recovery is defined in Section 5.2.2.2.

To categorize the well to sick progression during this period (from day=1 to the time before recovery), for each endpoint, the change from baseline will be summarized as follows:

a. The number and percentage of participants who have less than two days of available data for the endpoint during this period.

b. For each participant who has two or more days of available data for the endpoint during this period, a linear regression of the change from baseline vs. day, without intercept, will be fit. The participants without a statistically significant trend from the model will be categorized as non-responder, and number/percentage will be summarized.

c. For each participant with a statistically significant trend from the model in b, two measures will be computed: the day at which the first maximum change from baseline value (direction determined by linear fit in b) is reached during the period (defined as Tmax), and the ratio of the maximum change to the Tmax (defined as Growth Rate). The descriptive summary of the two measures for these significantly responding participants will be presented.

In addition, the well to sick progression period for each endpoint will be modelled monotonically. Firstly, a linear mixed effect model for the change from baseline vs. day will be used, with slope (i.e. without intercept) included as both fixed effects and random effects (to account for participant level variation). The number of participants used by the model, estimate of the slope's fixed effect mean, standard error, 90% CI, and p-value, as well as the variance estimate for its random effect will be presented.

Secondly, for the statistically significant endpoints determined by the linear mixed effect model, a logistic sigmoid growth/Emax model will also be considered for the population model:

$$y_{ij} - a_{i0} = \frac{d - a_{i0}}{1 + \left(\frac{day_j}{c}\right)^{-b}} + \varepsilon_{ij}$$

where $i$represents the ith participant, j represents the jth day, a_{i0} is the baseline value for the ith participant, and y_{ij} is the continuous endpoint value modelled as response. Parameters b, c, d represent the Hill slope, the day at which 50% effect is reached, and the asymptotic maximum sick response for the model, respectively, which can have participant level variation.

To facilitate model convergence, the maximum value observed during the period for each participant (direction determined by individual linear fit from b above), $\max_i$, will be used to substitute parameter d, and a linear mixed effect model of the logit transformed $\log((\max_i - y_{ij})/(y_{ij} - a_{i0}))$ vs log(day) will be used for the longitudinal changes, with its intercept and slope included as both fixed effects and random effects. The number of participants used by the model, estimates of the fixed effect means, standard errors, 90% CIs, and p-values, as well as the variance estimates for their random effects will be presented. The back-transformed mean estimates and 90%CIs for the parameters b and c, where $b = -\text{slope}$ and $c = \exp(-\text{intercept}/\text{slope})$, will also be presented.

Additional models may also be considered as appropriate, and the models' goodness of fit will be assessed using Akaike/Bayesian information criteria.

5.2.2.4. Well State Analysis

For each of the continuous endpoints during the well state, i.e. from enrollment day to the day before the first occurrence of new or increased symptoms in participants based on Electronic diary, the average of the endpoint values across all days will be summarized for the participants descriptively.

A mixed effects model will also be conducted, with participant and day as random effects, to evaluate the inter-participant variation and intra-participant day by day variation. Gender, age and race may be included as covariates to assess the effects. Estimates of the overall mean value per feature, its standard error, and 90% CI, as well as the variance components will be presented.

5.2.2.5. Correlation Between Symptoms and Voice Features

Change from baseline in symptom, after first occurrence of new or increased symptoms and before return to well state based on Electronic diary, will be modelled using a mixed-effects model with repeated measures (MMRM) with the change from baseline in the voice feature as covariate, day (categorical factor) as fixed effect, and participant as random effect. For the purpose of this analysis, the baseline is defined as the average of the endpoint values during the period up to 7 days before the first occurrence of new or increased symptoms, and the first occurrence of new or increased symptoms at day = 1. The day of return to well state is defined as the first day after the Maximal Score Day (defined in Section 5.2.2.2) when the total symptom score is less than or equal to the Initial Symptom Score (defined in Section 5.2.2.1).

Autoregressive AR(1) variance-covariance structure will be used for the MMRM. The estimate of the coefficient for the change in voice feature, its standard error, 90% CI, and p-value will be presented.

Pearson and Spearman rank order correlations between changes in symptoms and voice features will also be computed for each participant across days. These correlations will be summarized across participants. A heatmap plot will also be presented for each of these correlations.

CCI

[redacted]

CCI [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

5.2.3. Analyses for Categorical Endpoints

For Electronic diary symptom measures that have ordinal categorical responses, the values will be converted to numeric values (i.e. 0-4 for fever and 0-7 for other symptoms, where 0 = no symptoms), and analyzed as continuous endpoints according to Section 5.2.2.

5.2.4. Analyses for Time-to-Event Endpoints

N/A.

5.3. Methods to Manage Missing Data

All summaries and analyses will be based on observed data and missing data imputation is not planned.

6. ANALYSES AND SUMMARIES

Throughout the analyses, a participant is defined as symptomatic if they report at least one new or increased symptom in the Electronic diary during the course of the study.

6.1. Primary Endpoint(s)

Change in self-reported symptom scores in the Electronic diary from well to sick in symptomatic participants with RT-PCR confirmed SARS-CoV-2, influenza virus, or RSV; Change in voice features, such as pitch, jitter, harmonicity, entropy, flatness, shimmer from the voice collection as captured by the Electronic diary from well to sick in symptomatic participants with RT-PCR confirmed SARS-CoV-2, influenza virus, or RSV.

6.1.1. Main Analysis

6.1.1.1. Well to Sick

- Estimand: Primary Estimand E1(Section 2.1.1). This estimand is intended to provide a population level estimate of the well state as well as changes in symptoms and voice features from well to sick in symptomatic SARS-CoV-2, influenza virus, or RSV positive participants using Electronic diary.

- Analysis Sets: (a) Participants in Primary Analysis Set (Section 4) who become SARS-CoV-2, influenza virus or RSV positive (at Swab#2) and symptomatic during the study; (b) A subset of (a) who are symptom free at enrollment; (c) A subset of (a) who have symptoms due to chronic pre-existing conditions at enrollment. These analysis sets will be considered as deemed appropriate by data.

- Intercurrent events and missing data: (a) There are no particular intercurrent events considered. Participants who have events that lead to missing data will be included and missing data will not be imputed. (b) Inadequate compliance – data will be used as recorded.

- Statistical Method: Summary statistics and linear mixed effect analyses as described in Section 5.2.2.3. The total symptom score, each symptom and each voice feature will be modelled individually.

- Summaries:
  a. Descriptive summaries and corresponding plots of the mean/SD will be provided. For linear mixed effect model analyses, estimates for all parameters fixed effect means, standard errors, 90% CIs, and p-values, as well as the variance estimates for their

random effects will be provided (as described in Section 5.2.2.3). Results from all models will be presented in a table.

b. To assess the well-to-sick analysis baseline: For voice features that are determined as statistically significant based on linear model mixed effect analysis, the difference of the voice feature values between every pre-symptomatic day (starting from enrollment day) and the well-to-sick baseline (i.e. 7-day average before symptomatic onset as described in Section 5.2.2.3) will be computed. The mean/SD of the differences across participants in the analysis set will be plotted by day up to the symptomatic onset.

6.1.1.2. Well State

- Estimand: Primary Estimand E1 (Section 2.1.1).

- Analysis Sets: (a) Primary Analysis Set (Section 4). (b) A subset of (a) who are symptom free at enrollment; (c) A subset of (a) who have symptoms due to chronic conditions at enrollment. These analysis sets will be considered as deemed appropriate by data.

- Intercurrent events and missing data: (a) There are no particular intercurrent events considered. Participants who have events that lead to missing data will be included and missing data will not be imputed. (b) Inadequate compliance – data will be used as recorded.

- Statistical Method: Summary statistics and mixed effect analyses as described in Section 5.2.2.4. The total symptom score, each symptom and each voice feature will be modelled individually.

- Summaries: Descriptive summaries, and for mixed effect model estimates of the overall mean value per feature, its standard error, and 90% CI, as well as the variance components will be presented.

6.1.1.3. Correlation of Symptoms and Voice Features

- Estimand: Primary Estimand E1(Section 2.1.1).

- Analysis Sets: (a) Participants in Primary Analysis Set (Section 4) who become SARS-CoV-2, influenza virus, or RSV positive (at Swab#2) and symptomatic during the study; (b) A subset of (a) who are symptom free at enrollment; (c) A subset of (a) who have symptoms due to chronic conditions at enrollment. These analysis sets will be considered as deemed appropriate by data.

- Intercurrent events and missing data: (a) There are no particular intercurrent events considered. Participants who have events that lead to missing data will be included and missing data will not be imputed. (b) Inadequate compliance – data will be used as recorded.

- Statistical Method: MMRM model, and Pearson and Spearman rank order correlations, as described in Section 5.2.2.5.
- Summaries: Estimates of the coefficient for the change in voice features, standard errors, 90% CIs, and p-values will be summarized. Results from all models and the correlations will be presented in a table. Correlation plots will be presented.

6.1.2. Sensitivity/Supplementary Analyses

Sensitivity analyses may be conducted to assess the effect of compliance. Participants who have less than 3 out of 7 days of baseline data prior to symptom onset will be excluded and analyses in Section 6.1.1 repeated.

6.2. Secondary Endpoint(s)

Estimand: The estimand for all secondary endpoints is based on Estimate E2 (Section 2.1.2), which is intended to provide a population level estimate of the percentage of occurrence for the secondary endpoints.

Analysis Set: The analysis set for all secondary endpoints is based on FAS (as described in Section 4).

Intercurrent events and missing data: (a) There are no particular intercurrent events considered. Participants who have events that lead to missing data will be included and missing data will not be imputed. (b) Inadequate compliance – data will be used as recorded.

Statistical methods and summaries for each of the secondary endpoints are detailed below.

6.2.1. Percentage of total days of symptoms and total days of voice recordings entered in the Electronic diary

For each participant, compute:

a. Symptom compliance = number of compliant days* of symptoms entered in the Electronic diary divided by the total number of days in the study, and expressed as a percentage. Compute mean, SD, median, maximum and minimum percentage of days of symptoms across participants.

b. Voice compliance = number of compliant days* of completed voice recordings entered in the Electronic diary divided by the total number of days in the study, and expressed as a percentage. Compute mean, SD, median, maximum and minimum percentage of days of completed voice recordings across participants.

c. Partial voice compliance = number of compliant days* of partially completed voice recordings entered in the Electronic divided by the total number of days in the study, and expressed as a percentage. Compute mean, SD, median, maximum and minimum percentage of days of partially completed voice recordings across participants.

d. Symptom and voice compliance = number of days* with both symptom compliance and voice compliance, divided by the total number of days in the study, and expressed as a percentage. Compute mean, SD, median, maximum and minimum percentage of days of symptoms across participants.

*Since participants may complete the Electronic diary multiple times per day, a compliant day will be counted as valid and included in these analyses if participants have one or more sessions of an Electronic diary completed for symptoms (a. and d.) or completed/partially completed for voice recordings (b., c. and d.).

The analysis set for this analysis will be the FAS (as described in Section 4), excluding those participants who experienced technical issues due to study operational errors (e.g., received the wrong instructions for self-swabs).

Plots of number of compliant days of completed symptoms, completed voice recordings and partially completed voice recordings along day for all participants will be generated.

6.2.2. Percentage of quality voice recordings based on signal to noise ratio, sound duration and background noise measures

Compute number of days of quality voice recordings (i.e. recordings exceeding signal-to-noise ratio (SNR) and duration thresholds) per task from the total number of voice compliant days (see Section 6.2.1) per participant, and express as a percentage:

a. Percentage of voice recordings with P1 ‘ah’ (SNR > 20dB AND duration > 3 sec) OR (SNR <= 20 dB AND Noise Loud = ‘N’ AND duration > 3 sec)

b. Percentage of voice recordings with P2 ‘ee’ (SNR > 20dB AND duration > 3 sec) OR (SNR <= 20 dB AND Noise Loud = ‘N’ AND duration > 3 sec)

c. Percentage of voice recordings with P3 ‘mm’ (SNR > 20dB AND duration > 3 sec) OR (SNR <= 20 dB AND Noise Loud = ‘N’ AND duration > 3 sec)

d. Percentage of voice recordings with Reading Task (SNR > 20dB AND duration > 10 sec) OR (SNR <= 20 dB AND Noise Loud = ‘N’ AND duration > 10 sec)

e. Percentage of quality voice recordings

In the case of multiple Electronic diary sessions per day, the session with the highest quality voice recordings for that day will be used in this analysis.

Compute summary statistics (mean, SD, median, maximum and minimum percentage of days) of these percentages across participants.

6.2.3. Percentage of participants with SARS-CoV-2; and/or Influenza; and/or RSV RT-PCR based positivity in self-swabs

Compute the percentage of participants who:

(a) Had a valid self-swab #1 result that was positive for

  a. SARS-COV-2

  b. Flu A

c. Flu B
d. RSV
e. Any of the above (combined)

(b) Had a valid self-swab #2 that was positive for

a. SARS-COV-2
b. Flu A
c. Flu B
d. RSV
e. Any of the above (combined)

6.2.4. Feasibility of obtaining self-swabs

6.2.4.1. Percentage of participants administering the self-swab at self-swab #1 and self-swab #2

a. Number of participants who administered self-swab #1 divided by total number of participants in the study, reported as a percentage.

b. Number of participants who administered self-swab #2 divided by total number of participants in the study, reported as a percentage.

c. Number of participants who administered self-swab #1 and self-swab #2 divided by total number of participants in the study, reported as a percentage.

The analysis set for this analysis will be the FAS (as described in Section 4), excluding those participants who experienced technical issues due to study operational errors (e.g., received the wrong instructions for self-swabs).

6.2.4.2. Percentage of participants reporting symptoms in the Electronic diary, who have a self-swab collected at or around symptom onset

a. Number of days from symptom onset (reporting new or increased symptoms in the Electronic diary) to Record Swab #2 day in the Electronic diary averaged across participants who administered self-swab #2. Compute mean, SD, median, maximum and minimum days across participants.

b. Number of participants who administered self-swab #2 and reported new or increased symptoms in the Electronic diary before Week 6 divided by the number of participants who reported new or increased symptoms in the Electronic Diary before Week 6, reported as a percentage.

c. Number of participants who administered self-swab #2 and reported new or increased symptoms in the Electronic diary before Week 6 divided by the number of participants who administered self-swab #2 before Week 6, reported as a percentage.

6.2.4.3. Percentage of self-swabs with valid (positive or negative) or invalid (non-reportable due to technical or self-collection failures) results

a. Number of self-swabs #1 with **valid (positive or negative) result** for each of SARS-CoV-2/Flu A/Flu B/RSV divided by total number of self-swabs #1 received, reported as percentages.

b. Number of self-swabs #2 with **valid (positive or negative) result** for each of SARS-CoV-2/Flu A/Flu B/RSV divided by total number of self-swabs #2 received, reported as a percentage.

c. Total number of self-swabs (#1 and #2) with **valid (positive or negative) result** for each of SARS-CoV-2/Flu A/Flu B/RSV divided by total number of self-swabs (#1 and #2) received, reported as a percentage.

d. Total number of self-swabs (#1 and #2) with **NOT DONE** divided by total number of self-swabs (#1 and #2) reported in the Electronic diary (Record Swab), reported as a percentage.

CCI [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted] [redacted]

[redacted]	[redacted]

CCI [redacted]	[redacted]
[redacted]	[redacted]
[redacted]	

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

CCI [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

6.4. Subset Analyses

The study participants and analysis sets may be split in three acute viral respiratory illness groups (SARS-CoV-2, Influenza virus, and RSV) and the primary analysis detailed in Section 6.1.1 may be conducted on each group separately, if the n for each of the groups is deemed sufficient. Summary plot of well-to-sick changes by day in symptomatic positive participants may also by generated in four gender/age sub-groups (Male >60 y/o; Male <=60 y/o; Female >50 y/o; Female <=50 y/o) for each of the symptoms and statistically significant voice features (based on primary analysis) if data allows.

6.5. Baseline and Other Summaries and Analyses

6.5.1. Baseline Summaries

Demographic data and specific components of the enrollment questionnaire as listed in Section 3.4 will be summarized, and their descriptive statistics (n / percentages and/or mean, median, maximum and minimum statistics as appropriate) summarized and tabulated.

6.5.2. Study Conduct and Participant Disposition

Participants' evaluation, disposition and discontinuation will be summarized according to Pfizer standards.

6.5.3. Study Treatment Exposure

N/A.

6.5.4. Concomitant Medications and Nondrug Treatments

N/A.

6.6. Safety Summaries and Analyses

The safety analysis set is the FAS. The incidence and severity of adverse events (AEs)/research related injuries (RRIs) will be summarized according to Pfizer standards.

7. INTERIM ANALYSES

7.1. Introduction

An interim analysis (IA) #1 will be conducted once the first 1000 participants have completed the study to reassess study feasibility and enrollment requirements, including sample size.

An additional IA #2 may be conducted to evaluate the study primary and secondary objectives, and the algorithm development analyses detailed in the appendix SAP based on Cohort 1 data, consisting of approximately 9,100 participants who have been enrolled by 15 Feb 2022, prior to protocol amendment #2.

7.2. Interim Analyses and Summaries

7.2.1. IA #1

The list of study feasibility and compliance measures that will be summarized for the IA is:

- Number of participants enrolled
- Number of participants completed
- Number of participants without symptoms at enrollment (including or excluding chronic symptoms)
- Self-swab #1, number completed and by its outcome

- Self-swab #2, number completed before and after Day 42, and by its SARS-CoV-2 outcome (if after Day 42, number symptomatic participants during Day 42 – Day 56)
- Average % day Electronic diary completion, and before or after self-swab #2
- Average % quality voice Electronic diary entries
- Number of well to symptomatic sick (SARS-CoV-2) participants

All 1000 participants (the IA analysis set) will be included in the interim analysis based on observed cases, and missing data will not be imputed. To define the well to symptomatic sick participants for the IA, all following conditions need to be met to be included:

- Without symptoms at enrollment (two participant sets will be considered: including or excluding chronic symptoms)
- Negative self-swab #1 results for SARS-CoV-2, Influenza, and RSV
- Positive self-swab #2 result for SARS-CoV-2
- Reported new or increased symptoms during the study

Number and its percentage, or median/minimum/maximum will be summarized as appropriate. The summary will be produced for all participants, as well as by gender/age groups (i.e. Male >60 y/o; Male <=60 y/o; Female >50 y/o; Female <=50 y/o).

In addition, for the IA analysis set as well as for the well to symptomatic sick SARS-CoV-2 participant sets (including or excluding chronic symptoms at enrollment), the number of participants with quality Electronic diary completion will be plotted along day for all participants, and by gender/age groups.

7.2.2. IA #2

The analyses and summaries for this IA will include the primary and secondary analyses detailed in Section 6.1 and Section 6.2 respectively, and the algorithm development analyses detailed in appendix SAP (Section 8 Appendix A).

8. APPENDICES

Appendix A. X9001293 Algorithm Development Statistical Analysis Plan

Appendix B. List of Abbreviations

Abbreviation	Term
AcRIS	acute respiratory illness surveillance
ADAM	adaptive moment estimation
AE	adverse event
AR	autoregressive
AUC	area under the curve
AWS	amazon web services
CI	confidence interval
CNN	convolutional neural network
dB	decibel
EOS	end of study
FAS	full analysis set
FFT	fast Fourier transform
FN	false negatives
FP	false positives
IA	interim analysis
LGBM	light gradient boosting machine
CCI	[redacted]
MFCC	mel frequency cepstral coefficients
MMRM	mixed-effects model with repeated measures
N/A	not applicable
PCA	principal component analysis
PPS	predictive power score
RNN	recurrent neural networks
ResNET	residual neural network
RFECV	recursive feature elimination with cross validation
ROC	receiver operating characteristics
RRIs	research related injuries
RT-PCR	reverse transcription polymerase chain reaction
RSV	respiratory syncytial virus
SAE	serious adverse event
SAP	statistical analysis plan
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SD	standard deviation
SGD	stochastic gradient descent
SHAP	shapley additive explanations
SNR	signal-to-noise ratio
SVM	support vector machine
STFT	short-time Fourier transform
TN	true negatives

Abbreviation	**Term**
TP	true positives
t-SNE	t-distributed stochastic neighbor embedding
UMAP	uniform manifold approximation and projection
CCI	[redacted]
VGG	visual geometry group
WAV	waveform audio file format
XGBoost	extreme gradient boosting
y/o	years old

Protocol X9001293

Acute Respiratory Illness Surveillance (AcRIS) by Monitoring Voice and Illness Symptom Changes using a Mobile application in a Low-Interventional Decentralized Study

Algorithm Development Statistical Analysis Plan (adSAP)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This Appendix to the X9001293 SAP provides the detailed methodology for summary and statistical analyses related to the development of the screening algorithm for Symptomatic COVID-19 illness using voice features and symptoms data collected in the X9001293 study.

2. DEFINITIONS AND CONVENTIONS

The longitudinal nature of the data collected in X9001293 study (i.e., well and sick state data from the same participants) allows for a robust estimation of well and sick states and provides a controlled dataset for the training of the algorithm.

Definition of Total Symptom Score: the sum of all the symptom responses in the Electronic Diary converted to a numeric as described in Section 5.2.2.1. of the main SAP.

3. ANALYSIS SETS (POPULATIONS FOR ANALYSIS)

Data for all participants will be assessed to determine if participants meet the criteria for inclusion in each analysis population prior to releasing the database and classifications will be documented per standard operating procedures.

Algorithm Population	Description
Full Analysis Set (FAS) as per the Main SAP	All participants who sign the informed consent document and enroll in the study, as defined in the main SAP. *Qualified participants are males or females aged 18 years and older (or the minimum state-specific age of consent if > 18) at Screening visit, who are willing and able to comply with daily symptom and voice assessments on the Electronic Diary application and other study procedures including self-collection of nasal swabs, are expected to be available for the duration of the study, and are capable of giving informed consent.*
Primary Analysis Set	Participants in FAS who have a negative RT-PCR test (at swab #1) and no acute symptoms (as determined by the enrollment questionnaire), and who become positive for SARS-CoV-2, influenza virus, or RSV (at swab #2) and symptomatic during the study.
Secondary Analysis Set A	Participants in FAS who have a negative RT-PCR test (at swab #1) and a negative RT-PCR test (at swab #2) irrespective of symptoms.
Secondary Analysis Set B	Participants in FAS who have a positive SARS-CoV-2, influenza virus, or RSV RT-PCR test (at swab #1) and are symptomatic.
CCI [redacted]	[redacted]

For the purpose of algorithm development and testing, the sick and well states are defined for each of the analysis sets (when applicable).

For Primary Analysis Set

The **sick state** is defined by considering a window of Electronic Diary data around a positive RT-PCR test at swab #2.

To determine the start of the sick state window, consider the difference in symptoms between consecutive recordings. The start of the sick window will be a symptom increase within 3 days prior to swab 2. If there is no symptom increase within these 3 days, the start of the window will be on swab #2 day.

If the number of days between swab #1 and swab #2 is less than or equal to 3, the sick state window will start on swab #2 day.

The *end of the window* is after swab #2 and defined by either:

a. The last time the total symptom score is above 40% of the maximum total symptom score, or
b. The final available datapoint if the total symptom score stays above 40% of the maximum score.

If the total symptom remains below 40% of the maximum total symptom score after swab #2, the end of the window will be the swab #2 day.

The **well state** is defined by considering a window of Electronic Diary data around a negative RT-PCR test at swab #1, from 8 days before swab #1 up to 14 days before a positive swab #2. If there are less than 14 days between swab #1 and swab #2, the well state window ends on the day of swab #1.

For Secondary Analysis Set A

The **well state** is defined by considering an:

- Analysis time window from 8 days before swab #1 up to 8 days after swab #1.
- Analysis time window of 8 days up to and including swab #2.

For Secondary Analysis Set B

The **sick state** is defined by considering an analysis time window of 3 days before and 3 days after swab #1, including swab #1 day.

CCI

[redacted]

4. GENERAL METHODOLOGY AND CONVENTIONS

4.1. Hypotheses and Decision Rules

Given the algorithm's intended use as a screening tool in the general population not guided by a healthcare professional, we would define the success criteria for the algorithm as achieving a sensitivity on par with that of a rapid antigen test for SARS-CoV-2 infection as observed in real world studies, which is between 60% and 80%[1]. Hence, the success criteria for the algorithm will be an observed sensitivity of 70% or better.

The specificity of the algorithm will be assessed with a target minimum specificity of 60% as observed in other studies that use audio signals to detect SARS-CoV-2 infection [2]. The goal is to optimize the sensitivity of the algorithm as the primary performance measure without compromising the minimum specificity, since this algorithm is meant to be used as a screening tool (high sensitivity) and not a diagnostic tool.The reference rapid antigen test benchmarks compared to RT-PCR test for SARS-CoV-2 are as follows[1]: sensitivity ≥ 80% in symptomatic individuals and sensitivity ≥ 60% in asymptomatic individuals.

The probability that the observed sensitivity for symptomatic RT-PCR positive participants in AcRIS is greater than 70% with a total number of 100 cases is given in the following table for two assumed true values of expected sensitivity.

Table 1. Probability that observed sensitivity in symptomatic RT-PCR positive participants in AcRIS is greater than 70% with total of 100 cases.

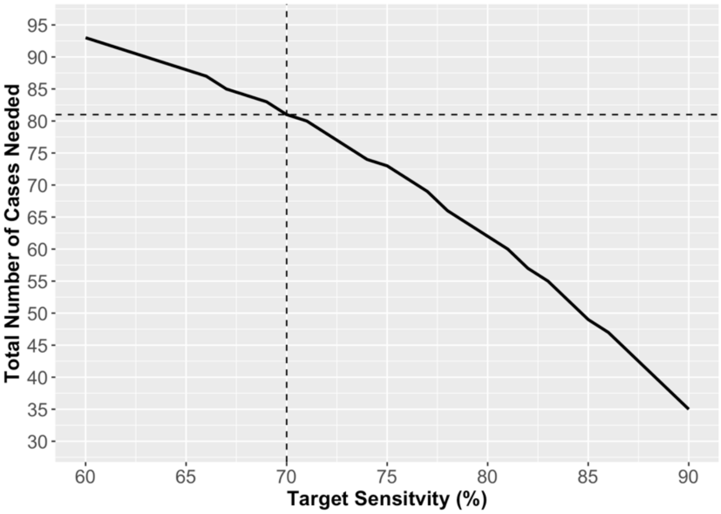
True Sensitivity	Probability of Observed Sensitivity > 70%[1]
60%	0.02
75%	0.88
80%	0.99

Thus, with at least 100 symptomatic RT-PCR positive participants in AcRIS, we hypothesize that the study has high probability of success under these conditions.

Furthermore, Figure 1. illustrates the number of cases required to achieve target sensitivity values for a given screening tool – e.g., for a sensitivity of 70% with 10% margin of error for the 95% confidence interval (i.e., 60% to 80%), 81 cases are required (assuming a binomial distribution for the sensitivity values). Identifying 100 symptomatic positive cases should allow us to attain 70% target sensitivity with precision of 8.98%.

[1] Under the assumption of a binomial distribution.

Figure 1. Estimation of the number of cases required to achieve target sensitivity values assuming a 10% margin of error.



4.2. General Methods

4.2.1. Voice Data Preparation

Audio data collected during the performance of voice tasks (i.e., phonemes and reading) in the Electronic Diary will be processed as follows:

Automatic checks are performed in the Electronic Diary for duration and background noise, with the latter used to drive the speech signal to a desired signal-to-noise ratio for analysis with feedback provided to the user so corrective action may be taken during data collection. In addition, cleaning will be performed on the audio data before data analysis to remove leading and trailing background noise associated with the speech signal. The data pipeline performing feature extraction will at a minimum have the following functions:

1. *Auto-cleaning*: to remove leading or trailing noise artifacts, log cleaning success or failure, and log percent change to audio signal.

2. *Automatic transcription:* The transcript and timing of each spoken word from the reading task will be automatically derived using proprietary Amazon Web Services (AWS) speech-to-text technologies [3]. Transcription content may be compared to the expected reading passage using the Levenshtein ratio metric[4] to measure differences between the transcribed speech and the canonical rainbow passage as a potential measure of incoming sample quality [5].

4.2.2. Voice Feature Extraction

The following list of features will be extracted from each voice task using established routines [6–8] for all participants in the FAS who have voice data and at least one RT-PCR test result. Each audio task will be matched with an appropriate subset of these features. Short-duration phoneme tasks EE and MM will be paired with features from lists 1, 2, and 3 below which generally focus on power, pitch, and spectral structure. Sustained phonation task AHH will be paired only with feature 4.d, which will provide information related to lung capacity. The reading task will be paired with features 3.f and 4.a-c, which cover both spectral structure and measures related to shortness of breath and breathlessness.

1. Features capturing power and power variability in speech:
    a. Shimmer
        i. Local
        ii. Local decibel (dB)
    b. Third octave band energy
2. Features capturing pitch and pitch variability in speech:
    a. Formants
        i. Mean formant values for formants 1, 2, 3.
        ii. Mean formant bandwidth values for formants 1, 2, 3.
    b. Jitter
        i. Local
        ii. Local absolute
    c. Coefficient of variation F0
    d. Voiced frames
3. Features capturing spectral structure of speech:
    a. Harmonicity
    b. Entropy
    c. Spectral flatness
    d. Voiced low-to-high ratio
    e. Cepstral peak prominence
    f. Mel-frequency cepstral coefficients
        i. Mean coefficients 1-13
        ii. Standard deviation coefficients 1-13
        iii. Delta coefficients 1-13
        iv. Delta-delta coefficients 1-13
4. Features capturing information related to shortness of breath, breathlessness, and/or lung capacity.
    a. Speaking rate
    b. Number of pauses
    c. Average pause length
    d. Maximum phonation time

4.2.3. Feature Selection and Dimensionality Reduction Methods

The univariate analysis performed as part of the primary objective of the study and detailed in the main SAP will be employed on the features listed in Section 4.2.2. This analysis will help inform which voice features and symptoms change with SARS-CoV-2 infection to guide the classical machine learning analysis described in Section 4.2.6. To minimize the amount of noise or irrelevant information in model training, validation, and testing, both feature selection and dimensionality reduction methods will be used. This process will include feature values paired with their respective sick and well labels (supervised learning) and may also include the symptom scores associated with each sample. This analysis will be performed on participants' data within the primary analysis set.

Clustering methods such as t-distributed Stochastic Neighbor Embedding (t-SNE) [9], Uniform Manifold Approximation and Projection (UMAP) [10], and/or 2-component Principal Component Analysis (PCA) [11] will be used to qualitatively investigate to what degree features and/or the combination of features and symptoms can separate the well and sick states. While the resulting clusters will not be used to directly remove certain features from consideration, they will help inform which features and/or symptoms group together, which will be useful in understanding how the final model makes predictions.

Feature and/or symptom correlations will provide an understanding of symmetrical linear relationships in the data. Features that are highly correlated with one another will be identified as features that could potentially be dropped, as they provide similar information. The Predictive Power Score (PPS)[12], which compares a feature's ability to predict another feature against the performance of a naïve classifier (e.g., choosing the most frequent class, choosing the median), will be used to investigate asymmetrical non-linear relationships among the features. Features with a PPS indicating low predictive utility (e.g., less than 0.1 using F1-score as an evaluation metric) will also be separated as features that can potentially be eliminated. Other metrics may also be used to assist in this analysis, such as the chi-square test, information gain, or random forest feature importance.

Once this process has been completed, automated feature elimination algorithms will be used iteratively to further remove any remaining redundant features. This will be accomplished with Recursive Feature Elimination with Cross Validation (RFECV) [13], which recursively removes features until predictive performance degrades, and/or the Boruta algorithm [14], which removes features that fail to perform better than a random classifier.

Finally, PCA may be used to perform dimensionality reduction by projecting the remaining features from a high-dimensional space m to a lower-dimensional space n. Depending on the size of the remaining feature set, if PCA can significantly reduce the number of features (e.g., n < 75% of the total number of features m) while maintaining a high degree of explained variance (e.g., >95%), then it may be appropriately used to reduce the final feature set. Assuming PCA is deemed to be a viable option, both the resulting PCA features, that is, those n components that explain a high percent (e.g, >95%) of the variance, *and* the features remaining after RFECV/Boruta will be used as separate inputs for performing the repeated stratified k-fold cross-validation outlined in Section 4.2.6.1. This way, the cross-validation performances of the PCA feature set and the reduced feature set can be compared to ensure that this last dimensionality reduction step does not significantly impact performance.

4.2.4. Spectrogram and Mel-spectrogram Extraction

Spectrograms of each of these WAV (waveform audio file format) audio files will be computed to obtain a visual representation of the spectrum of audio frequencies against time [15]. The spectrogram will be log-transformed, converting it from an amplitude representation to a decibel representation. Spectrograms are generated by performing the Short-Time Fourier Transform (STFT) with a sliding Hamming or Hann window. Examples of parameters are sliding window size 32ms, hop window of 10ms, and a 512-point Fast Fourier Transform (FFT). This would give spectrograms of size 257 x 201 for 2 seconds of audio data recorded with a 16k Hz sampling rate. The resulting spectrogram will be integrated into Mel-spaced frequency bins (such as 64), where the magnitude of each bin is log transformed [16]. The Mel spectrogram has the advantage over a typical frequency spectrogram because the Mel scale in the Mel spectrogram has unequal spacing in the frequency bands and provides a higher resolution in lower frequencies, as compared to the equally spaced frequency bands in a normal spectrogram.

The Mel spectrogram follows speech recognition systems. The lower coefficients which map to the lower frequencies associated with human hearing capabilities will be examined. There are several methods for converting the frequency scale to the Mel. One way to convert frequency *f* into Mel scale m as [17]:

$$m = 2595 \times log_{10}(1 + \frac{f}{700})$$

Cepstral analysis will be performed on the Mel spectrum of audio samples to compute their Cepstral coefficients, so called Mel Frequency Cepstral Coefficients (MFCC)17. The extracted MFCC features result in a two-dimensional matrix for every sample, where each column represents one signal frame, and each row represents extracted MFCC features for a specific frame. MFCC features can be used in several ways for classification. Spectrograms, Mel spectrograms and MFCC will be used as input to the Deep Learning Modeling described in Section 4.2.6 18,[19].

4.2.5. Classical Machine Learning

Using the reduced feature set from Section 4.2.3., classical machine learning algorithms will be trained to perform the binary classification of well versus sick states. Ensemble learning algorithms are the preferred method because they: typically generalize well, are very robust to overfitting, can be trained efficiently, require little hyper-parameter tuning to achieve competitive results, perform predictions quickly and can be easily packaged for deployment [20]. The classical machine learning approach also lends itself to making understandable predictions. SHapley Additive exPlanations (SHAP)[21] will be used to investigate the most successful models to determine to what degree each feature influences model output. The main classical machine learning models of interest include: Random Forest [22]; Bagging [23]; Extreme Gradient Boosting (XGBoost)[24] ; Light Gradient Boosting Machine (LGBM) [25]; Support Vector Machine (SVM) Ensemble [26].

Additional modeling may include a combination of spectrograms (described in Section 4.2.4.) and voice features described in Section 4.2.2. The spectrogram averaged across time may be combined with the voice features and used as the input to the classical machine learning model. In early fusion, a classifier is trained on the combined set of voice features and averaged spectrograms. In late fusion, separate models are trained on averaged spectrograms and voice features, and the final prediction is the average of the output from the two models [27–29].

In addition, covariance approach which quantifies the linear relationship between features will also be explored to perform the binary classification of well versus sick states. This approach has been shown to be invariant to the linear transformation of the features[30]. This invariance property might diminish the intersubject variability of the acoustic features and can also reduce any learning effects of the acoustic features over time on the model performance. The input data for this approach will be the MFCCs matrix, as described in Section 4.2.4.

Specifically, the covariance matrix between MFCCs will be estimated using the Ledoit-Wolf shrinkage estimator.[31] Each covariance matrix is an instance of symmetric positive definite matrix and represents a point on the Riemannian manifold.[32] To apply classical distance-based machine learning algorithms, the resulting covariance matrix can be mapped to the tangent space[33] which is an instance of a vector space. This transformation to the tangent space also permits the fusion of other acoustic features in the covariance model. Finally, classical machine learning models like Random Forest [22]; Balanced Random Forest[34], Bagging [23]; Extreme Gradient Boosting (XGBoost)[24] ; Light Gradient Boosting Machine (LGBM) [25]; Support Vector Machine (SVM) Ensemble [26] will be explored in the tangent vector space to perform the binary classification of well versus sick states.

4.2.5.1. Classical Machine Learning Model Training and Testing

The data will be randomly divided into two sets by subject: training (e.g., 80% of subjects) and test set (e.g., 20% of subjects). Repeated stratified k-fold cross validation (on the subject level) on the training data will be used to evaluate performance, as it provides confidence in the resulting performance values. Both k (number of folds) and n (number of repetitions) will be chosen based on the final sample size. The test set will be a hold-out test set until model training and validation is complete.

4.2.5.2. Classical Machine Learning Model Selection

In order to evaluate model performance, predictions from each model will be summarized in a confusion matrix, as shown in Table 2. Performance metrics including *sensitivity*, *specificity* and *balanced accuracy* will be computed as shown in Table 3. Area under the receiver operating characteristic curve (AUC-ROC) showing *sensitivity* vs. *1 – specificity* will also be assessed. A proposed classical machine learning model will be chosen based on the repeated stratified k-fold cross validation performance. This model may be further tuned to maximize performance, will be trained on the entire training set to generate a final model, and will finally be tested on the hold-out test set to generate the final performance results.

Table 2. Confusion Matrix. True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN); in terms of total number among predictions.

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Table 3. Definition of Performance Measures. True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN).

Performance Measure	Formula
Sensitivity	TP / (TP + FN)
Specificity	TN / (TN + FP)
Balanced Accuracy	(Sensitivity + Specificity) / 2

4.2.5.3. Classical Machine Learning Model Output

The final output of the classical machine learning approach will be the full set of model predictions for the best-performing model's repeated k-fold cross-validation, as well as the full set of predictions of this model on the hold-out test set. Both sets of predictions will be paired with their respective labels so that all performance metrics can be calculated.

4.2.6. Deep Learning

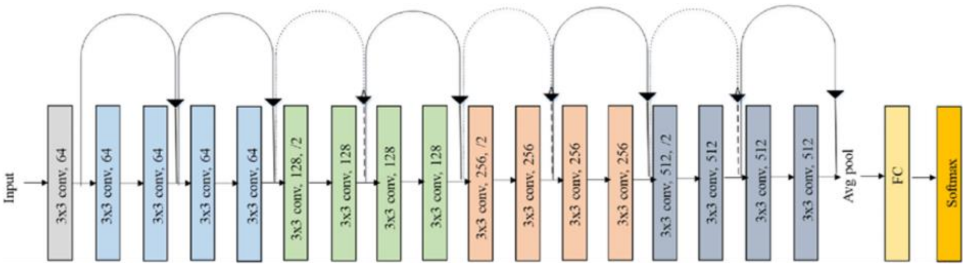
In addition to classical machine learning algorithms, deep learning models will be used to perform the binary classification of well versus sick states. The input data for the deep learning models will include spectrograms, Mel-spectrograms, or other variants of spectrogram such as log-Mel-spectrogram or MFCC, as described in Section 4.2.4 for the CNN models or the audio files in case of Transformer models.

Deep learning models have been widely used to classify covid positive and negative participants using voice or cough data (e.g., see review papers by Deshpande et al [35]and Qian et al,[36]). Most studies have used various architectures of convolutional neural network (CNN) models, and few have applied recurrent neural networks (RNN) [37]. More recently, Transformers have been found useful for speech analysis [38,39]. We may apply various architectures of the CNN models such as ResNet18 (Residual Neural Network 18) [40], VGG16 (Visual Geometry Group 16) [41], VGG19 [42] or custom built models.

We will choose the CNN architecures based on the dimension of the final dataset. The details of the CNN architectures that may be computed are as follows:

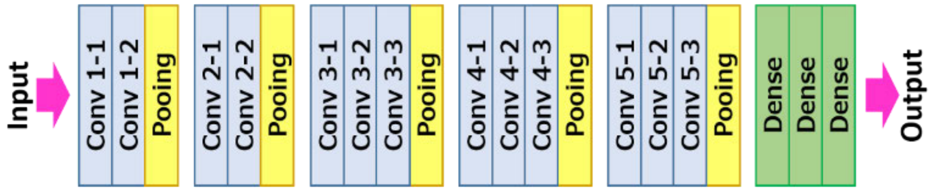
ResNet18: ResNet-18 is a convolutional neural network that is 18 layers deep (Figure 2). ResNet is more recent architecture than original CNN architectures like VGG. The advantage of ResNet is the skip connections, or shortcuts to jump over some layers. One of the problems of deep neural networks is that the first layers' weights will not change much during the training process. This is called vanishing gradient. The skip connections solve the problem of vanishing gradients in deep neural networks by allowing this alternate shortcut path for the gradient to flow through. These connections also help by allowing the model to learn the identity functions which ensure that a higher layer will perform at least as well as the lower layer, and not worse. Layers may be added or removed based on our data. Cross-validation will be used to determine if the model is overfitting the training data. In that case, layers will be removed to reduce the complexity of the model. If the model is underfitting the data, more layers will be added.

Figure 2. ResNet18 architecture. Conv stands for convolution; FC is fully-connected layer and SoftMax is the activation function used in the last layer. The last layer will be adjusted using a sigmoid function for a binary classification. Figure source: Ramzan et al., 2019[43].



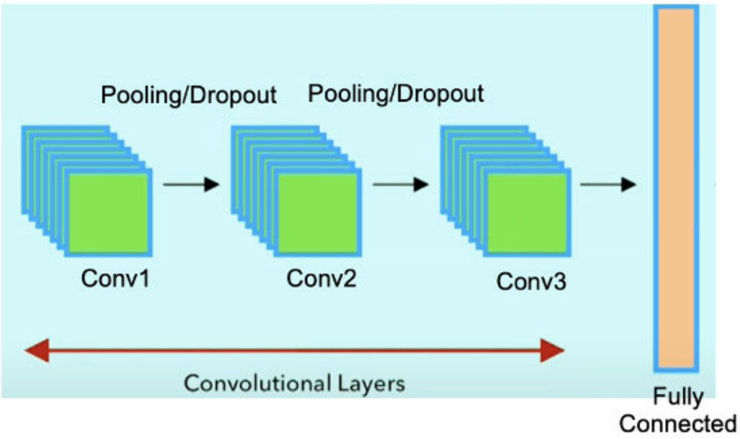
VGG16: VGG16 is a simple and widely used CNN architecture with 16 layers depth (Figure 3), which was originally built for ImageNet, a large visual database project used in visual object recognition software research. VGG16 is popular due to its ease of implementation and is used in many deep learning image classification techniques. Recently, VGG16 and VGG19 has been used for classifying SARS-CoV-2 infection positive and negative individuals using voice data [42,44]. Layers may be added or removed based on our data.

Figure 3. VGG16 architecture. Figure source: https://neurohive.io/en/popular-networks/vgg16/



State-of-the-art architectures: Studies [19,45,46] using state-of-the-art CNN architectures which are less dense (<10 layers) than typical standard CNN models (>16 layers), have been shown to be successful for SARS-CoV-2 classification as well. Therefore, CNN architectures that do not follow standard architectures (e.g., ResNet or VGG) will be built. This allows for the choosing/adjustment of the number of layers, kernel size, filter size, padding, stride, pooling type and size, batch normalization, and number of nodes in the fully connected layer based on the classification task. Dropout will be used at various layers for regularizing the parameters to help prevent the neural network from overfitting [47]. Lastly, we might add Long Short-Team Memory layer to the CNN model which will help to find temporal relationships [48]. A schematic display of a conventional CNN architecture is shown in Figure 4.

Figure 4. Example CNN model with three convolutional layers followed by fully connected layer. The last layer is the classification layer which has only two nodes with sigmoid function as the activation function. Figure adapted from Kulchyk et al, 2019 [49]



In addition to the above described modeling, we may try two open source, Transformers-based acoustic models that use audio files as the input.

The first model is the Nvidia Connectionist Temporal Classification model[50], an encoder-decoder based transformer model originally developed as a speech-to-text model trained on thousands of hours of human speech. The pretrained encoder model will be used to generate acoustic features and develop a classification algorithm to predict whether a participant is sick or well.

The second model is Wav2vec2[38] wav, a self-supervised, deep-learning transformer model which learns latent encodings of audio recordings using contrastive predictive coding tasks. A custom wav2vec2 model will be trained on the reading task and a deep-learning classification model will be trained to predict whether a participant is sick or well.

4.2.6.1. Pre-Training

Studies have shown that pre-training significantly improves the performance of the model. For example, Bagad et al. [40] found 17% improvement in area under the curve (AUC) when using a pre-trained model compared to one without pre-training. This is particularly important when dealing with small (<1000 samples) or medium size data (<10,000 samples). Therefore, we may use this approach. For example, one way of pre-training the model is to classify “mm” vs “ah” using the data from participants that were negative at both swab #1 and swab #2. Using transfer learning the weights of this model can be used for a sick vs. well classifier [19]. Other pre-training approaches will be considered.

4.2.6.2. Data Augmentation

A common approach to increase the sample size or balance the data when dealing with imbalanced datasets is to use data augmentation. It has been shown that data augmentation helps to improve the performance of the model [18,51]. Various data augmentation methods may be used such as adding gaussian noise to the audio file, pitch shifting, shifting the time signal and stretching the time signal [18,52].

4.2.6.3. Deep Learning Model Training and Testing

The dataset will be divided into three separate datasets: training (e.g., 70% of data), validation (e.g., 10% of data) and test set (e.g., 20% of data). We will replicate the training and validation split three times at the subject level to ensure the results are not biased by single randomization. The training of the model requires the choice of an error function, so called a loss function, that can be used to estimate the loss of the model so that the parameters can be updated to reduce the loss on the next evaluation. Given that we have a binary classification problem (sick vs well), our loss function would be binary cross entropy, or weighted binary cross entropy. To optimize the parameters of the model, we will either use ADAM or Stochastic Gradient Descent (SGD) approach. The name ADAM is derived from adaptive moment estimation. ADAM uses momentum of the optimization object and adaptive learning rate to converge faster than SGD[53]. We will use learning rates such as 0.001 or 0.0001 as the starting point.

4.2.6.4. Deep Learning Model Selection

To evaluate the performance of the deep learning models, performance metrics including *sensitivity*, *specificity* and *balanced accuracy* will be derived on the test and validation sets (Table 2 and Table 3). The model *AUC-ROC* will be assessed. The final model will be trained on the entire training set to maximize fine tuning, and will finally be tested on the hold-out test set to generate the final performance results.

4.2.6.5. Deep Learning Model Output

The final output of the deep learning approach will be the full set of model predictions for the best-performing model validation, as well as the full set of predictions of the final model on the hold-out test set for the three replications. Both sets of predictions will be paired with their respective labels so that all performance metrics can be calculated.

4.3. Fusion of Classical Machine Learning and Deep Learning Modeling

Methods such as stacking[54] or bagging[23] may be used to combine the outputs of the classical machine learning and deep learning models. Performance of the combined models will be assessed as detailed in Section 4.2.6.4.

4.4. Well vs. Sick Analysis with Sick State Definition Centered Around Maximum Total Symptom Score

This analysis will investigate the distinction of well and sick states from voice features, with the sick state definition centered around the maximum total symptom score instead of centered around a positive RT-PCR test. For the purpose of this analysis, the sick and well states will be defined as follows:

For participants who have a negative RT-PCR test at swab #1 and a positive RT-PCR test at swab #2:

The **sick state** is defined by considering a window of Electronic Diary data centered around the maximum total symptom score, from 3 days before to 3 days after the maximum total symptom score, with the following conditions:

- If there are less than or equal to 3 days between maximum total symptom score and swab #1, the start of the sick period will be the day of the maximum total symptom score.

- If there are more than 3 days between the maximum total symptom score and swab #1, the start of the sick period must be at least 3 days after swab #1.

- If there are multiple days with maximum total symptom score, consider the one closest to the day of swab #2 as the center of the window.

The **well state** is defined by considering a window of Electronic Diary data around a negative RT-PCR test at swab #1, from 8 days before swab #1 up to 14 days before the maximum total symptom score. If there are less than 14 days between swab #1 and the maximum total symptom score, the well state window ends on the day of swab #1.

For participants who have a positive RT-PCR test at swab #1, the sick state definition is as follows:

The **sick state** is defined by considering a window of Electronic Diary data centered around the maximum total symptom score, from 3 days before to 3 days after the maximum total symptom score.

- If there are multiple days with maximum total symptom score, consider the one closest to the day of swab #1 as the center of the window.

For participants who have a negative RT-PCR test at both swab #1 and swab #2, the well state definition will remain as defined in Section 3, namely:

The **well state** is defined by considering an:

- Analysis time window from 8 days before swab #1 up to 8 days after swab #1
- Analysis time window of 8 days up to and including swab #2.

This well and sick data will serve as the input to the machine learning and deep learning analysis, and the models will be trained, validated and tested as described in Section 4.2.5. and Section 4.2.6.

4.5. Change from Baseline Analysis

This analysis will investigate change from individual-level baseline in participants who start the study as well (negative swab #1) and who become sick and symptomatic during the study (positive swab #2). The individual-level baseline will be defined by up to the first three available days of that participant's data within the first 7 days in the study.

The individual baseline will be subtracted from endpoint values that fall within the sick and well state windows as defined in Section 3. This data will serve as the input to the machine learning models. The models will be trained, validated and tested as described in Section 4.2.5. Deep learning modeling may also be performed on this data as described in Section 4.2.6.

For the covariance model described in Section 4.2.5, MFCC features will be used as the input to the model. The MFCCs matrices are not aligned in time across Electronic Diary sessions. Moreover, the difference between two positive definite matrices is not positive definite in general [55]. To account for this, MFCCs from the baseline will be mapped to the tangent space using methods described in Section 4.2.5. Then the projected baseline vector will be subtracted from well or sick vectors in this tangent space.

4.6. Further Model Testing

The best-performing output models from the classical machine learning (Section 4.2.5.) deep learning (Section 4.2.6.), fusion method (Sections 4.3.) and well vs. sick analysis with sick state definition centered around maximum symptom score (Section 4.4.) will be tested on data of:

(a) Negative cases to get an estimate of the false positive rate.

(b) Symptomatic positive participants at swab #1 as an additional dataset for testing the sensitivity rate of the models.

CCI [redacted]

For the Change From Baseline Analysis (Section 4.5.), the best performing model will be tested on participants who are negative at both swab #1 and swab #2 to investigate the false positive rate of this approach.

CCI [redacted]

[redacted]

CCI [redacted]

CCI

CCI

4.8. Methods to Manage Missing Data

All summaries and analyses will be based on observed data and missing data imputation is not planned. Inadequate compliance data will be used as recorded.

5. ANALYSES AND SUMMARIES

5.1. Classification of Well and Sick States using Classical Machine Learning Models

- Analysis Set: Participants in Primary Analysis Set (Section 3).

- Statistical Method: Classical machine learning modeling will be performed as described in Section 4.2.5. The best performing model will be selected based on performance metrics listed in Section 4.2.5.2.

- Summaries:

Using the output of the best performing model described in Section 4.2.5.3. the following results will be generated for the specified analysis set (data permitting):

  - Listing of the features used in the final model
  - Summary statistics of the performance measures listed in Section 4.2.5.2. (mean and standard deviation) for classification of sick and well states for the validation set. Results will be presented in a table.
  - The confusion matrix for the test set will be tabulated and a listing of the performance measures provided.

5.2. Classification of Well and Sick States Using Deep Learning Models

- Analysis Set: Participants in Primary Analysis Set if deemed appropriate by data.

- Statistical Method: Deep learning modeling will be performed as described in Section 4.2.6. The best performing model will be selected based on performance metrics listed in Section 4.2.6.4.

- Summaries:

Using the output of the best performing model described in Section 4.2.6.5., the following results will be generated for the specified analysis set (data permitting):

  - Listings of the specifications of the spectrograms and mel-spectrograms used for the final deep learning model (Section 4.2.4.)
  - Summary statistics of the performance measures listed in Section 4.2.6.4. (mean and standard deviation) for classification of sick and well states for the validation set. Results will be presented in a table.
  - The confusion matrix for the test set will be tabulated and a listing of the performance measures provided.

If a fusion model is deemed appropariate by data (as described in Section 4.3.), the following will be generated:

  - Listings of the voice features (Section 4.2.2.) and specifications of the spectrograms and mel-spectrograms used for the final model (Section 4.2.4.)
  - Summary statistics of the performance measures listed in Section 4.2.6.4. (mean and standard deviation) for classification of sick and well states for the validation set. Results will be presented in a table.
  - The confusion matrix for the test set will be tabulated and a listing of the performance measures provided.

5.3. Well vs. Sick Analysis with Sick State Definition Centered Around Maximum Total Symptom Score

- Analysis Set: Participants in Primary Analysis Set (Section 3) if deemed appropriate by data.

- Statistical Method: Listed in Section 4.4.

- Summaries:

Using the output of the best performing classical machine learning (Section 4.2.5.), deep learning (Section 4.2.6.) or fusion method (Section 4.3.), the following results will be generated for the specified analysis set (data permitting):

  - Listing of the features used in the final classical machine learning model and/or Listings of the specifications of the spectrograms and mel-spectrograms used for the final deep learning model (Section 4.2.4.)
  - Summary statistics of the performance measures listed in Section 4.2.5.2. (mean and standard deviation) for classification of sick and well states for the validation set. Results will be presented in a table.
  - The confusion matrix for the test set will be tabulated and a listing of the performance measures provided.

5.4. Change from Baseline Analysis

- Analysis Set: Participants in Primary Analysis Set (Section 3).

- Statistical Method: Listed in Section 4.5.

- Summaries: Using the output of the best performing classical machine learning (Section 4.2.5.), deep learning (Section 4.2.6.) or fusion method (Section 4.3.) , the following results will be generated for the specified analysis set:
    - Listing of the features used in the final model
    - Summary statistics of the performance measures listed in Section 4.2.6.2. (mean and standard deviation) for classification of sick and well states for the validation set. Results will be presented in a table.
    - The confusion matrix for the test set will be tabulated and a listing of the performance measures provided.

5.5. Further Model Testing

5.5.1. Model Testing on RT-PCR Negative Participants

- Analysis Set: Participants in Secondary Analysis Set A to test models trained on the Primary Analysis Set (Section 3).

- Statistical Method: (a) As listed in Section 4.6., the best-performing output model from the classical machine learning (Section 4.2.5.), deep learning (Section 4.2.6.), fusion method (Section 4.3.) and well vs. sick analysis with sick state definition centered around maximum total symptom score method (Section 4.4.**)** will be tested on negative cases to estimate the false positive rate of the models. (b) The best performing Change From Baseline model will also be tested on analysis set above.

- Summaries: The following results will be generated for the specified analysis set (data permitting):
    - Confusion matrix will be tabulated and a listing of the performance measures provided for the two models (a) and (b).
    - Table of values of *specificity* as a function of Total Symptom Score thresholds for both models.
    - Plot of *specificity* against Total Symptom Score thresholds for both models.

5.5.2. Model Testing on Positive Symptomatic Participants at Swab #1

- Analysis Set: Participants in Secondary Analysis Set B (Section 3) to test model trained on the Primary Analysis Set.

- Statistical Method: As listed in Section 4.6., the best-performing output model from the classical machine learning (Section 4.2.5.), deep learning (Section 4.2.6.), fusion method (Section 4.3.) and well vs. sick analysis with sick state definition centered around maximum total symptom score method (Section 4.4.) will be tested on participants who are positive and symptomatic at swab #1 as an independent test set.

- Summaries:
    - Confusion matrix will be tabulated and a listing of the performance measures provided.
    - Table of values of *sensitivity* as a function of Total Symptom Score thresholds.
    - Plot of *sensitivity* against Total Symptom Score thresholds.

CCI

[redacted]

CCI

[redacted]

5.7. Sensitivity/Supplementary Analyses

Additional analysis may be conducted to investigate the influence of potential covariates such as age, gender, race, and smoking status on the performance of the models. If data allows, tables with model performance measures (*sensitivity*, *specificity and balanced accuracy*) for different cohorts defined by these covariates will be generated.

6. REFERENCES

1. Schuit E, Veldhuijzen IK, Venekamp RP, et al. Diagnostic accuracy of rapid antigen tests in asymptomatic and presymptomatic close contacts of individuals with confirmed SARS-CoV-2 infection: cross sectional study. *BMJ*. July 2021:n1676. doi:10.1136/bmj.n1676

2. Han J, Xia T, Spathis D, et al. Sounds of COVID-19: exploring realistic performance of audio-based digital testing. *npj Digit Med*. 2022;5(1):16. doi:10.1038/s41746-021-00553-x

3. AWS A. Transcribe. https://docs.aws.amazon.com/transcribe/.

4. Levenshtein VI, others. Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet Physics Doklady*. Vol 10. ; 1966:707-710.

5. Fairbanks G. Voice and Articulation Drillbook. *Laryngoscope*. 1941;51(12):1141. doi:10.1288/00005537-194112000-00007

6. McFee B, Raffel C, Liang D, et al. librosa: Audio and Music Signal Analysis in Python. In: ; 2015.

7. Jadoul Y, Thompson B, de Boer B. Introducing Parselmouth: A Python interface to Praat. *J Phon*. 2018;71:1-15. doi:10.1016/j.wocn.2018.07.001

8. Boersma P, Weenink D. Praat: doing phonetics by computer. 2021.

9. der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9(11).

10. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw*. 2018;3(29):861. doi:10.21105/joss.00861

11. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos Mag*. 1900;5(50 (302)):157–175.

12. Wetschoreck F, Krabel T, Krishnamurthy S. 8080labs/ppscore: zenodo release. 2020.

13. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46(1):389-422.

14. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *J Stat Softw*. 2010;36(11). doi:10.18637/jss.v036.i11

15. Sejdić E, Djurović I, Jiang J. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digit Signal Process*. 2009;19(1):153-183. doi:10.1016/j.dsp.2007.12.004

16. Stevens SS, Volkmann J, Newman EB. A Scale for the Measurement of the

Psychological Magnitude Pitch. *J Acoust Soc Am*. 1937;8(3):185-190. doi:10.1121/1.1915893

17. Patel K, Prasad RK. Speech recognition and verification using MFCC \& VQ. *Int J Emerg Sci Eng(IJESE)*. 2013;1(7):137-140.

18. Brown C, Chauhan J, Grammenos A, et al. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. June 2020. doi:10.1145/3394486.3412865

19. Imran A, Posokhova I, Qureshi HN, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Informatics Med Unlocked*. 2020;20:100378. doi:10.1016/j.imu.2020.100378

20. Sagi O, Rokach L. Ensemble learning: A survey. *WIREs Data Min Knowl Discov*. 2018;8(4). doi:10.1002/widm.1249

21. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9

22. Ho TK. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol 1. ; 1995:278-282.

23. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140. doi:10.1007/BF00058655

24. Chen T, Guestrin C. XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM; 2016:785-794. doi:10.1145/2939672.2939785

25. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146-3154.

26. Wandekoken ED, Varejão FM, Batista R, Rauber TW. Support Vector Machine Ensemble Based on Feature and Hyperparameter Variation for Real-World Machine Fault Diagnosis. In: ; 2011:271-282. doi:10.1007/978-3-642-20505-7_24

27. Huang S-C, Pareek A, Zamanian R, Banerjee I, Lungren MP. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Sci Rep*. 2020;10(1):22147. doi:10.1038/s41598-020-78888-w

28. Lahat D, Adali T, Jutten C. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc IEEE*. 2015;103(9):1449-1477. doi:10.1109/JPROC.2015.2460697

29. Khaleghi B, Khamis A, Karray FO, Razavi SN. Multisensor data fusion: A review of

the state-of-the-art. *Inf Fusion*. 2013;14(1):28-44. doi:10.1016/j.inffus.2011.08.001

30. Porikli FM, Tuzel O, Meer P. Covariance Tracking using Model Update Based on Means on Riemannian Manifolds. In: ; 2005.

31. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *J Multivar Anal*. 2004;88(2):365-411. doi:10.1016/S0047-259X(03)00096-4

32. Förstner W, Moonen B. A Metric for Covariance Matrices. In: *Geodesy-The Challenge of the 3rd Millennium*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003:299-309. doi:10.1007/978-3-662-05296-9_31

33. Barachant A, Bonnet S, Congedo M, Jutten C. Multiclass Brain–Computer Interface Classification by Riemannian Geometry. *IEEE Trans Biomed Eng*. 2012;59(4):920-928. doi:10.1109/TBME.2011.2172210

34. Chen C. Using Random Forest to Learn Imbalanced Data. In: ; 2004.

35. Deshpande G, Batliner A, Schuller BW. AI-Based human audio processing for COVID-19: A comprehensive overview. *Pattern Recognit*. 2022;122:108289. doi:10.1016/j.patcog.2021.108289

36. Qian K, Schuller BW, Yamamoto Y. Recent Advances in Computer Audition for Diagnosing COVID-19: An Overview. In: *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE; 2021:181-182. doi:10.1109/LifeTech52111.2021.9391791

37. 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;1:268-274. doi:10.1109/OJEMB.2020.3026468

38. Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. June 2020.

39. Chi P-H, Chung P-H, Wu T-H, et al. Audio ALBERT: A Lite BERT for Self-supervised Learning of Audio Representation. May 2020.

40. Bagad P, Dalmia A, Doshi J, et al. Cough Against COVID: Evidence of COVID-19 Signature in Cough Sounds. September 2020.

41. Mohammed EA, Keyhani M, Sanati-Nezhad A, Hejazi SH, Far BH. An ensemble learning approach to digital corona virus preliminary screening from cough sounds. *Sci Rep*. 2021;11(1):15404. doi:10.1038/s41598-021-95042-2

42. Suppakitjanusant P, Sungkanuparph S, Wongsinin T, et al. Identifying individuals with recent COVID-19 through voice classification using deep learning. *Sci Rep*. 2021;11(1):19149. doi:10.1038/s41598-021-98742-x

43. Ramzan F, Khan MUG, Rehmat A, et al. A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer’s Disease Stages Using Resting-State fMRI and Residual Neural Networks. *J Med Syst*. 2020;44(2):37. doi:10.1007/s10916-019-1475-2

44. Nweke HF, Teh YW, Al-garadi MA, Alo UR. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Syst Appl*. 2018;105:233-261. doi:10.1016/j.eswa.2018.03.056

45. Lella KK, Pja A. Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: Cough, voice, and breath. *Alexandria Eng J*. 2022;61(2):1319-1334. doi:10.1016/j.aej.2021.06.024

46. Aly M, Rahouma KH, Ramzy SM. Pay attention to the speech: COVID-19 diagnosis using machine learning and crowdsourced respiratory and speech recordings. *Alexandria Eng J*. 2022;61(5):3487-3500. doi:10.1016/j.aej.2021.08.070

47. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(1):1929-1958.

48. Singh VP, Kumar S, Jha RS. *Samsung R and D Bangalore DiCOVA 2021 Challenge System Report*.; 2021.

49. Kulchyk J, Etemad A. Activity Recognition with Wearable Accelerometers using Deep Convolutional Neural Network and the Effect of Sensor Placement. In: *2019 IEEE SENSORS*. IEEE; 2019:1-4. doi:10.1109/SENSORS43011.2019.8956668

50. Gulati A, Qin J, Chiu C-C, et al. Conformer: Convolution-augmented Transformer for Speech Recognition. May 2020.

51. Park DS, Chan W, Zhang Y, et al. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. April 2019. doi:10.21437/Interspeech.2019-2680

52. Fakhry A, Jiang X, Xiao J, Chaudhari G, Han A, Khanzada A. Virufy: A Multi-Branch Deep Learning Network for Automated Detection of COVID-19. March 2021.

53. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. December 2014.

54. Wolpert DH. Stacked generalization. *Neural Networks*. 1992;5(2):241-259. doi:10.1016/S0893-6080(05)80023-1

55. Bonnabel S, Sepulchre R. Riemannian Metric and Geometric Mean for Positive Semidefinite Matrices of Fixed Rank. *SIAM J Matrix Anal Appl*. 2010;31(3):1055-1070. doi:10.1137/080731347