Statistical Analysis Plan (SAP)

Title	Pilot Validation of a Prototype Mobile Health Tympanometer
CRU/Department/Division/Center	COM/Otolaryngology
IRB Number	274552
Investigators:	
Primary Investigator	Susan D. Emmett, MD, MPH
Collaborative Lead	Samantha Kleindienst Robler, AuD, PhD
Co-authors (if known)	
Analysis Biostatistician(s)	Milan Bimali, PhD
Subject Matter Expert	Samantha Kleindienst Robler, AuD, PhD Susan D. Emmett, MD, MPH
Original Creation Date	02/03/2023
Version Date	02/03/2023
Finalization Date	04/15/2024
Project Folder Location	TBD
Effort Estimate (optional)	

Investigator Agreement

☐ All statistical analyses included in an abstract or manuscript should reflect the work of the biostatistician(s) listed on this SAP. No changes or additional analyses should be made to the results or findings without discussing with the project biostatistician(s).

☐ All biostatisticians on this SAP should be given sufficient time to review the full presentation, abstract, manuscript, or grant and be included as co-authors on any abstract or manuscript resulting from the analyses.

☐ If substantial additional analysis is necessary or the aims of the project change, a new SAP will need to be developed.

☐ Publications resulting from this SAP are supported in part by the TRI award number PTC 2022-03 and be submitted to PubMed Central.

☐I have reviewed the SAP and understand that any changes must be documented.

Acknowledged by: Click or tap here to enter text.
Date: Click or tap to enter a date.

Activity Log

Acronyms

AC	Arkansas Children's
CI	Confidence interval
CNE	Could not evaluate
ECV	Ear Canal Volume
FDA	United States Food and Drug Administration
IRB	Institutional Review Board
ITT	Intent-to-Treat
IQR	Interquartile range
LOA	Limits of Agreement
LMICs	Low- and middle-income countries
mHealth	Mobile Health
ML	Machine Learning
PP	Per Protocol
REDCap	Research Electronic Data Capture
SA	Static Admittance
SAP	Statistical Analysis Plan
SD	Standard deviation
TPP	Tympanometric Peak Pressure
UAMS	University of Arkansas for Medical Sciences

Target Journal(s):
ClinicialTrials.gov number/link:

Contents

1 Study Overview

1.1 Background/Introduction:

The goal of the “Statistical Analysis Plan” (SAP) outlined here is to provide a comprehensive document that provides required details for the summary, visualization, and analysis of the data that is measured and/or observed during the course of the study "Pilot Validation of a Prototype Mobile Health Tympanometer". This SAP is finalized based on the study protocol finalized on 03/07/2024.
The protocol should be read with the understanding that the outlined methods related to summarizing, displaying, and analyzing the study data should be considered flexible, and deviations from the pre-planned approach may be required. Statistical analyses rely on satisfactorily meeting different assumptions that can be validated only during data analysis. Hence deviations from the pre-planned analysis approach can be inevitable. A statistical and/or clinical description justifying the need for these deviations will be included.

An estimated 1.6 billion people are living with hearing loss globally, making hearing loss the second leading impairment worldwide. Unfortunately, over 80% of affected individuals reside in low- and middle-income countries (LMICs) with limited access to hearing care. To be effective, screening must account for the type and prevalence of hearing loss in a given population. Most screening programs only use pure-tone screening and are not equipped to identify middle ear disease that is widespread in populations with a high prevalence of infection-related hearing loss. A major reason for this gap is that tympanometry, the device used to clinically identify middle ear disease, is not typically used for screening because it is expensive and designed to be operated by trained professionals. To address the barriers of cost and training, the study team has developed a low-cost, lay-friendly mobile health (mHealth) tympanometer and a novel machine learning (ML) algorithm that together simplify detection of middle-ear disease and interpretation of results. The new device has the potential to transform hearing screening in LMICs, where the burden of hearing loss is greatest and is not addressed by current hearing screening methodology.

To prepare the new device for large-scale testing by lay-screeners, the device needs to be validated. This study represents a first-in-human pilot validation with audiologists that will compare the performance of the prototype to a commercially available tympanometer in 60 ears (approximately 20 adults and 10 children). This study will also evaluate the performance of the ML algorithm using data from the prototype tympanometer. This early feasibility study of the mHealth tympanometer will be conducted in partnership with audiologists at the Audiology Clinics of the University of Arkansas for Medical Sciences (UAMS) and Arkansas Children’s (AC). Information from this study will guide design modifications in the device for large-scale validation studies needed to bring this evidence-based technology to underserved communities in the rural US and abroad.

References:
1. Protocol: Pilot Validation of a Prototype Mobile Health Tympanometer
2. R21-33 Proposal: A digital Innovation to Address Preventable Childhood Hearing Loss in Low- and Middle-income Countries

1.2 Study Aim

This study will evaluate the accuracy and performance of the prototype device compared to commercial tympanometer in pediatric and adult population.

1.3 Research Objectives and Study Hypotheses

Objective 1 (primary):

To obtain preliminary estimates of the accuracy of the prototype device compared to commercial tympanometer based on audiologist's interpretation (for both devices).

Hypothesis 1:

The diagnostic accuracy (sensitivity and specificity) of the prototype device based on audiologist's interpretation as compared to the commercial tympanometer based on audiologist's interpretation is 80% or greater.

Objective 2 (primary):
To compare categorical classifications of tympanogram types (A, B, C) based on audiologist's interpretation compared to the machine learning algorithm's interpretation.

Objective 2A:
To obtain preliminary estimates of the agreement between audiologist's interpretation of the prototype device and machine learning algorithm's interpretation of the prototype device.

Hypothesis 2A:
The categorical classifications of tympanogram types based on audiologist's interpretation of the prototype device versus machine learning algorithm's interpretation of the prototype device are not significantly different.

Objective 2B:
To obtain preliminary estimates of the accuracy of the prototype device based on machine learning algorithm's interpretation compared to commercial tympanometer based on audiologist's interpretation.

Hypothesis 2B:

The diagnostic accuracy (sensitivity and specificity) of the prototype device based on machine learning algorithm's interpretation as compared to the commercial tympanometer based on audiologist's interpretation is 80% or greater.

Objective 3 (secondary):
To investigate the differences in numerical measures (ear canal volume, static admittance and tympanometric peak pressure) between the commercial and prototype devices.

Hypothesis 3:
There are no statistically significant differences in numerical measures (ear canal volume, static admittance and tympanometric peak pressure) between the commercial and prototype devices.

Objective 4 (secondary):
To compare audiologists' perceptions on ease of use between commercial and prototype devices based on the 4-item 5-point Likert scale survey from audiologists.

Hypothesis 4:
There are no statistically significant differences between commercial device and prototype device ease of use.

2 Study Methods

2.1 Study Design

Pilot Validation: This is a pilot validation study involving a comparative investigation of a commercially available tympanometer with a minimal risk lay-friendly (mHealth) tympanometer prototype. The pilot validation will be conducted in consented adults and children presenting to the Audiology Clinics of UAMS and AC. The study will evaluate device performance compared to a commercial tympanometer in patients with various middle ear pathologies. Data acquired from the pilot validation will also be used with our machine learning algorithm to determine how well it works with prototype data and the if there will be a need for further refinement.

Eligible patients will be invited to participate, and the consent form reviewed. Consented participants will receive their audiological services as scheduled, with the addition of prototype measurements to their appointment. The audiology appointment will include routine ear and hearing measures, such as otoscopy (visual ear exam) and pure tone testing. The prototype device will be conducted before the commercial tympanometry. Audiologists will be asked to complete the prototype measurement and interpretation prior to the commercial device to limit bias in study results.
The resulting tympanograms from the prototype device and the clinical tympanometer will be interpreted by the audiologist and entered in a secure Research Electronic Data Capture (REDCap) database. The audiologist will also answer a few brief questions after testing each participant in a REDCap form on device performance for both the prototype and commercial tympanometers.
Specifically, the order of events will be as follows:

1. Audiologist will test each ear of the participant with the prototype device and interpret the tympanogram. These findings will be entered in a secure Research Electronic Data Capture (REDCap) database in real time using a tablet or computer.
2. Next, the audiologist will test each ear of the participant with a commercial tympanometer, interpret the results and enter them in the database as described above.
3. Finally, the audiologist will complete 4 questions on user experience with the prototype and commercial devices and document it in a secure REDCap database.

Testing the machine learning algorithm: The performance of the machine learning algorithm will be tested by using raw data from the prototype mHealth tympanometer obtained from the pilot validation study. The ML will interpret the raw tympanometric data and classify the results into Type A, B, and C. The classification by the ML algorithm will be compared to audiologists' interpretation, and concordance will be assessed. The audiologists interpreting the data will be blind to the ML algorithm, and the ML algorithm interpretation will not be provided the audiologist.

2.2 Power and Sample size

A sample size of 30 participants (60 ears), with an assumed prevalence of 0.20 will yield a half-width of 95% confidence interval ranging from 0.25-0.33 when the sample sensitivity is 0.80; and a half-width of 95% confidence interval ranging from 0.12-0.18 when the sample specificity is 0.80. The range of the half-width of 95% confidence interval reflect the extreme scenario of the intra-class correlation ranging from 0 to 1.

2.3 Sampling Strategy

Convenience sampling will be employed for this study. We will recruit a minimum of 30 participants, which will include adults (n=20) presenting to the Audiology Clinic of University of Arkansas for Medical

Sciences (UAMS) and pediatric participants (n=10) presenting to Arkansas Children's (AC) Audiology Clinic. In total, a minimum of 60 ears will be included in the analysis.

2.4 Study Population

Adult and pediatric patients presenting to audiology clinics of UAMS and AC with various middle ear pathologies.

2.5 Inclusion Criteria

Audiology Patient Participants

- Individuals, 1-year and older
- Presenting to the UAMS or AC Audiology Clinics for evaluation where tympanometry is warranted for testing at the discretion of the audiologist (current practice).
- Presence of various middle ear health states/pathologies that result in Type A, B, C tympanograms; examples include normal, occluding cerumen, effusion, perforation, retraction otosclerosis, cholesteatoma, ossicular chain discontinuity, myringitis, and tympanosclerosis
- English-speaking

2.6 Exclusion Criteria

- Children or adults with cognitive disabilities
- Unable to provide consent/assent
- Individuals who are unable to sit still
- Any other condition, that, in the opinion of the investigator, might interfere with the safe conduct of the study or place the participant at increased risk

2.7 Data Acquisition

Table 1. Data acquisition

Study design	Pilot validation study
Data source/how the data were collected	REDCap Survey: *Link will be added later* Data collected during study visit: • Enrollment form • Data collection form ° Audiologist Otoscopy Interpretation ° Audiologist Interpretation and Performance for prototype device ° Audiologist Interpretation and Performance for commercial tympanometer • ML Algorithm Interpretation form
Contact information for team member responsible for data collection/acquisition	
Date or version (if downloaded, provide date)	
Data transfer method and date	
Where dataset is stored	REDCap

3 Analysis Population

The analysis population will consist of all subjects who are enrolled in the study and on whom both the study devices have been administered.

4 Outcomes, Exposures, and Additional Variables of Interest

4.1 Outcome(s)

Table 2. Description of outcome variables

Outcome	Description	Variables and Source	Specifications
Primary Outcome(s)			
Categorical classification of tympanogram types A, B, C 1) Audiologist's interpretation of commercial device versus audiologist's interpretation of prototype device 2A) Audiologist's interpretation of prototype device versus machine learning algorithm's interpretation of prototype device 2B) Audiologist's interpretation of commercial device versus machine learning algorithm's interpretation of prototype device	Tympanograms can be classified into types (Type A, B, C) that help to determine if an eardrum has fluid behind it, is retracted, or perhaps has a perforation, all common ear disease states. Type A indicates normal ear function, and types B and C indicate abnormal middle ear function.	Audiologist commercial device classification (left, right): Tymp_type_au_com_l; Tymp_type_au_com_r Audiologist prototype device classification (left, right): Tymp_type_au_proto_l; Tymp_type_au_proto_r Machine learning algorithm prototype device classification (left, right): Tymp_type_ml_proto_l; Tymp_type_ml_proto_r Forms: 1) Data collection form: Audiologist interpretation and performance for mHealth tympanometer 2) Data collection form: Audiologist interpretation and performance for commercial tympanometer 3) ML Algorithm Interpretation form	1=A; 2=B; 3=C; 4=CNE (could not evaluate)
Secondary Outcome(s)			
Ear canal volume, static admittance and	Numerical measures	**Commercial** device (left, right):	Numerical measures

tympanometric peak pressure	Commercial device versus prototype device	ecv_com_l; sa_com_l; tpp_com_l; ecv_com_r; sa_com_r; tpp_com_r **Prototype** device (left, right): ecv_proto_l; sa_proto_l; tpp_proto_l; ecv_proto_r; sa_proto_r; tpp_proto_r Forms: 1) Data collection form: Audiologist Interpretation and Performance for commercial tympanometer 2) ML Algorithm Interpretation form	
Device performance	Device performance will be evaluated using a 4-item 5-point Likert scale survey	Prototype device performance: perform_proto_Q1 perform_proto_Q2 perform_proto_Q3 perform_proto_Q4 Commercial device performance: perform_com_Q1 perform_com_Q2 perform_com_Q3 perform_com_Q4 Forms: 1) Data collection form: Audiologist Interpretation and Performance for mHealth tympanometer 2) Data collection form: Audiologist Interpretation and Performance for	1=Completely disagree; 2=Disagree; 3= Neither Agree nor Disagree; 4=Agree; 5=Completely Agree

		commercial tympanometer	

4.2 Additional Variables of Interest

Table 3. Description of additional variables

Variable	**Description**	**Variables and Source**	**Specifications**
Age (years)	Age in years	ag_yr Form: Enrollment form	Range from 1-89
Age group	Age groups	Ag_group Form: Enrollment form	Participants with Age >89 will be categorized as one group 1=1-89; 2=>89
Sex	Sex of adult or child	sex Form: Enrollment form	1=Male; 2=Female; 3=Transgender Female; 4=Transgender Male; 5=Gender Variant / Non-Conforming/Not listed/Prefer not to Answer
Race	Race of adult or child	Race_American Indian or Alaska Native; Race_Asian; Race_Black or African American; Race_Native Hawaiian or Other Pacific Islander; Race_White; Race_Other; Race_Other, specify; Race_Unknown or Prefer Not to Answer Form: Enrollment form	1=Yes; 0=No
Ethnicity	Ethnicity of adult or child	Ethnicity Form: Enrollment form	1=Hispanic or Latino; 2=Not Hispanic or Latino; 3=Unknown or Prefer Not to Answer
Location	Location of the recruitment	Location Form:	1=UAMS; 2=AC

		Enrollment form	
Otoscopy findings by ear		Oto_left Oto_right Data Collection Form: Audiologist Interpretation for Otoscopy	0=Normal; 1= Non-occluding cerumen; 2= Occluding cerumen; 3= Retraction; 4= Effusion; 5= Acute otitis media; 6= Otorrhea; 7= Perforation; 8= Patent tube; 9= Plugged tube; 10= External otitis; 11= Myringitis; 12= Foreign body; 13=Cholesteatoma 14=Tympanosclerosis 15= Could not evaluate; 16= Other

5 Analysis Plan

5.1 General Considerations

Analysis timeline: The final statistical analysis will be performed after study data is available from 30 participants (60 ears); and the Center for Hearing Health Equity data management team has completed the transfer of an analysis-ready dataset to the study statistician.

Statistical Software: The analysis will be done using SAS (version 9.4) and R (version 4.3.2).
A two-sided alpha of 0.05 will be used to determine statistical significance.

5.2 Data Summary and Visualization

Descriptive statistics will be presented as n (non-missing sample size); mean (standard deviation) and median (inter quartile range) for continuous variables and as frequency (percentage, based on non-missing sample size) for categorical variables. Descriptive statistics will be presented as overall and by study site (UAMS – adult population vs AC – pediatric population).

5.3 Analysis of Primary Endpoints

The primary outcome is the classification of the tympanogram types, which can be categorized into one of the three categories A, B, and C. The classification of tympanogram types from the commercial device (based on audiologist's interpretation) will be considered as reference standard (gold standard). In the subsequent section, the unit of cluster will refer to subjects who contribute to tympanogram types for left and right ears using the prototype as well as the commercial device. The analysis will be done based on a 2-level outcome, namely A (normal) vs B/C (diseased).

Diagnostic accuracy of the prototype device based on audiologist's interpretation: The diagnostic accuracy of the prototype device (based on audiologist's interpretation) will be estimated using sensitivity and specificity. The sensitivity and specificity will be estimated using a binary outcome – A (normal) vs B/C (diseased). To ensure the maximum use of the available data, the estimates of sensitivity and specificity will

be computed at ear-level. Taking into consideration the potential correlation between the two ears, the sensitivity and specificity will be estimated using generalized estimating equation (GEE), with a robust (sandwich) variance estimator and unstructured correlation matrix.

Secondary analysis of primary endpoints: The secondary analysis of primary endpoints will constitute additional measures of diagnostic accuracy namely – positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (PLR), and negative likelihood ratio (NLR). To obtain estimates of PPV and NPV, the disease prevalence will be set at 0.20 (estimate of disease prevalence may be updated).

Analysis based on 3-level outcome: The performance of the prototype device will be presented using a three-by-three confusion matrix. The overall accuracy will be computed from the confusion matrix. We will also compute separate estimates of sensitivity and specificity separately for tympanometric tracing types B and C respectively.

Agreement between audiologist's and machine learning algorithm's interpretation of the tympanometric tracing types (based on prototype device): The agreement between audiologist's and machine learning algorithm's interpretation of the tympanometric tracing (based on prototype device) will be tested for statistical significance using Durkalski's adjustment to the McNemar's test statistics for the analysis of clustered matched-pair data (Durkalski et al). Durkalski's method for the analysis of clustered matched-pair data will maximize the use of available data and adjusts for multiple units within a cluster while avoiding any assumptions related to intra-cluster correlation. The Kappa statistic and the corresponding 95% confidence interval will be estimated using cluster-bootstrapped approach (Kang et al). The bootstrapped approach will ensure maximal data usage while also taking into account the potential subject-level clustering. The aforementioned measures of agreement will be based on a 3-level outcome.

Diagnostic accuracy of the prototype device based on machine learning algorithm's interpretation: The diagnostic accuracy of the prototype device (based on machine learning algorithm's interpretation) will be estimated using sensitivity and specificity. These estimates will be obtained using the analysis approach described earlier for diagnostic accuracy of prototype device based on audiologist's interpretation. The secondary analysis of diagnostic accuracy as well as the analysis based on 3-level outcome will be done using the approach described earlier for the prototype device based on audiologist's interpretation.

5.4 Analysis of Secondary Endpoints

Comparison of ear canal volume (ECV), static admittance (SA) and tympanometric peak pressure (TPP): The agreement in ECV, SA, and TPP between the commercial device and prototype device using Bland-Altman approach. The difference plot (plot of difference in measurement vs mean measurement) will be used to visualize the difference as well as to assess the assumption of normality for the differences. The limits of agreements (LoA) will be constructed: mean of difference $\pm$ 1.96×SD of difference. It is expected that LoA will capture 95% of the differences in measurements between the two devices. The distributional assumptions related to Bland-Altman method will be validated (e.g. normality of difference will be tested for statistical significance using Shapiro-Wilk's test) and alternate approaches such as data transformation may be employed if needed. It is anticipated that the difference in ECV and TPP between the commercial device and prototype device will be within 0.4 and 40 measurement units respectively. The difference in distribution of the ECV, SA, and TPP will be assessed (separately) for statistical significance using Wilcoxon signed rank test.

Comparison of ease of device usage: The ease of device usage based on audiologist's assessment will be measured using a 4-question survey. Each survey question is presented on 5-point Likert scale (Completely Disagree, Disagree, Neither agree or disagree, Agree Completely, Agree). The response will be assigned a score of 1-5 (corresponding to Completely Disagree-Completely Agree) with higher score indicating greater

degree of agreement. For each audiologist an aggregated score will be obtained. The aggregated score will range from 4-20 with higher score suggesting greater ease of use. Each audiologist will contribute two assessments – one for the prototype device and the other for the commercial device. The difference in aggregated score between the two devices will be tested for statistical significance using Wilcoxon signed rank test.

5.5 Subgroup analysis

The descriptive statistics and the analysis related to primary endpoints will be presented separately for adults and pediatric populations. Owing to small sample size, these results will be considered exploratory.

5.6 Safety Analysis

Adverse events (AE) constitute any untoward or unfavorable occurrence in a research participant associated with the participant's involvement in the research that may or may not be related to the individual's participation in research. The adverse events will be collected and coded in accordance with the guidelines specified in MedDRA/CTCAE.

Serious adverse events (SAE) refer to the adverse events that meet at least one of the following criteria:

- Leads to death
- Considered life threatening (putting participants in immediate risk of death)
- Necessitates inpatient hospitalization or prolongation (if already hospitalized)
- Results in persistent or significant disability/incapacity
- Requires medical or surgical intervention to prevent outcomes listed in the above definition of SAE
- Causes significant psychological, social, economic, or legal harm to participants or others
- Results in breach of confidentiality that is damaging to participants' rights, employment, financial standing or reputation

Adverse events will be analysed and listed as per the guidelines listed below:

1. If a participant experiences the same AE multiple times during the study, the patient will be counted only once in the number of participants experiencing the event.
2. If a patient experiences the same AE multiple times but with different severity during the study, the worst or most intense event will be counted.
3. If there are adverse events that are not coded, the summary table will use the exact description from as reported in the database. The statistician may consult with the study investigator for further information to resolve this (i.e. whether it can be reassigned to one of the existing coding)

The safety analysis will be descriptive, and no inferential statistics will be done. Summary statistics will be presented by severity and device relatedness.

5.7 Multiple testing

No adjustment for multiple testing is planned for this pilot study.

5.8 Sensitivity analysis

N/A

5.9 Missing data

There could be missing values from either prototype or commercial device (e.g., Table 4, Ear 1, 2, 3 or Ear 5, 6) or both the devices (e.g., Ear 4). The analysis for all study objectives will be based on complete case analysis (the last column of Table 4 i.e., data available for both devices).

Missing data in the outcome variable arises from inability to take tympanometry measurements for a participant (CNE status), especially for children with the presence of ear pathology, or younger children less likely to sit still for measurements. We will not include an ear in the main analysis if data from one device

(either commercial or prototype) or both devices are missing for that ear. No multiple imputation is planned for this study.

The frequency and proportion of missing data will be presented separately for the prototype and the commercial device. Missing proportions will be compared for commercial and prototype devices by ear status, adults/children, number of ears and tested for statistical significance using either Chi-square test or Fisher's exact test.

Table 4: Example of missing values by device type

	Commercial device	Prototype device	Complete case analysis for all study objectives
Ear 1	√	missing	
Ear 2	√	missing	
Ear 3	√	missing	
Ear 4	missing	missing	
Ear 5	missing	√	
Ear 6	missing	√	
Ear 7	√	√	√
Ear 8	√	√	√
Ear 9	√	√	√
Ear 10	√	√	√
…Ear 60			

6 Limitations

Missing data are more likely to occur for ears that have presence of pathology, because it is potentially more difficult to obtain a reading from a diseased ear. We may also find missingness for children who have higher rates of ear disease and can be more difficult to test. We may over or underestimate the level of agreement between the two devices especially for the ears with pathology.

7 Appendix

7.1 Shell Tables and Figures

The table, listing, figure shells are provided in a separate document.