OPEN

# Hypothesis Tests for Continuous Audiometric Threshold Data

Zechen Liu,[1,6] Zhuoran Wei,[1,6] Jiaxuan Li,[1] Gary Curhan,[2,3,4,5] Sharon Curhan,[2,3,4,7] and Molin Wang[1,2,3,4,7]

**Objectives:** Hypothesis tests for hearing threshold data may be challenging due to the special structure of the response variable, which consists of the measurements from the participant's two ears at multiple frequencies. The commonly-used methods may have inflated type I error rates for the global test that examines whether exposure-hearing threshold associations exist in at least one of the frequencies. We propose using both-ear methods, including all frequencies in the same model for hypothesis testing.

**Design:** We compared the both-ear method to commonly used single-ear methods, such as the worse-ear, better-ear, left/right-ear, average-ear methods, and both-ear methods that evaluate individual audiometric frequencies in separate models, through both theoretical consideration and a simulation study. Differences between the methods were illustrated using hypothesis tests for the associations between the Dietary Approaches to Stop Hypertension adherence score and 3-year change in hearing thresholds among participants in the Conservation of Hearing Study.

**Results:** We found that (1) in the absence of ear-level confounders, the better-ear, worse-ear and left/right-ear methods have less power for frequency-specific tests and for the global test; (2) in the presence of ear-level confounders, the better-ear and worse-ear methods are invalid, and the left/right-ear and average-ear methods have less power, with the power loss in the left/right-ear much greater than the average-ear method, for frequency-specific tests and for the global test; and (3) the both-ear method with separate analyses for individual frequencies is invalid for the global test.

**Conclusions:** For hypothesis testing to evaluate whether there are significant associations between an exposure of interest and audiometric hearing threshold measurements, the both-ear method that includes all frequencies in the same model is the recommended analytic approach.

**Key words:** Audiometry, Bias, Data correlation, Generalized estimating equation, Hearing loss, Hypothesis test, Pure-tone audiometry.

**Abbreviations:** AE = average-ear; BE = better ear; CHEARS = Conservation of Hearing Study; DASH = Dietary Approaches to Stop Hypertension; GEE = generalized estimating equation; PTA = pure-tone average; WE = worse ear.

(Ear & Hearing 2024;45;1165–1172)

## INTRODUCTION

Hearing loss was a leading cause of disability in 2015; over 5% of the world's population suffers from disabling hearing loss (Wilson et al. 2017; Chadha et al. 2021; WHO 2021). Moreover, the adverse influence of hearing loss on health and quality of life is considerable (Mick et al. 2014; Dawes et al. 2015), thus research to identify potentially modifiable risk factors for hearing loss that could inform strategies for prevention is a pressing public health priority (Chadha et al. 2021; Haile et al. 2021).

In epidemiological studies of hearing health, researchers are often interested in assessing whether associations exist between a given exposure and the hearing threshold measures. However, statistical methods for the evaluation of exposure-hearing associations using audiometric threshold data are inconsistent (Cruickshanks et al. 2003; Bainbridge et al. 2008). Researchers commonly used a single-ear (e.g., the worse-ear [WE], better-ear [BE], left/right-ear) hearing measurement as the response variable (Verschuur et al. 2012; Grondin et al. 2015; Lin et al. 2016) or analyze the data for each frequency or group of frequencies separately (Hu et al. 2018; Shih et al. 2020). Definitions of the single-ear method may vary and are typically based on the pure-tone average (PTA) of threshold measurements at a prespecified set of frequencies. In a longitudinal setting, the outcome could be the frequency-specific PTA threshold measurements at each follow-up time point, or alternatively the change in frequency-specific PTA at each postbaseline follow-up time point. The WE method and BE method are based on measurements of the WE, the ear with a higher threshold or change in threshold, and the BE, the ear with a lower threshold or change in threshold, respectively. Table 1 shows an example of the WE and BE outcomes for the low-frequency PTA threshold measurements. In the longitudinal study setting, the ear which is the WE or BE could change over time. For example, in Table 1, the WE is the left-ear at visit 1 but the right-ear at visit 2.

Existing publications demonstrate the performance of different estimation methods (e.g., bias and efficiency of WE/BE, average-ear [AE], and both-ear methods) for continuous outcomes (Sheng et al. 2022) and binary outcomes (Chen et al. 2022). This article focuses on continuous outcomes. For continuous outcome scenarios, when there are only participant-level confounders, using WE or BE methods leads to unbiased but less efficient estimators; lower efficiency means the estimators have larger variance (Sheng et al. 2022). If the information from only one ear at each time point is used in the analyses, then the information from the other ear is ignored. Therefore, the

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA; [2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA; [3]Harvard Medical School, Boston, Massachusetts, USA; [4]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA; and [5]Renal Division, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA; [6]These authors contributed equally as co-first authors; [7]These authors contributed contributed equally as co-senior authors.



Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and text of this article on the journal's Web site (www.ear-hearing.com).





<zdoi; 10.1097/AUD.0000000000001503>

**TABLE 1. Example of PTA of threshold data for low frequencies (0.5, 1, 2 kHz) from 1 participant in a hearing loss study**

| | Single-Ear Method | | | | | |
|---|---|---|---|---|---|---|
| Visit | Right | Left | WE Outcome | BE Outcome | AE Outcome | Both-Ear Outcome |
| Visit 1 (baseline) | 15 | 10 | 15 | 10 | 12.5 | (15, 10) |
| Visit 2 | 15 | 20 | 20 | 15 | 17.5 | (15, 20) |
| Change | 0 | 10 | 10 | 0 | 5 | (0, 10) |

*Data in this table are fabricated.*
*AE, average-ear; BE, better-ear; Left, left-ear; PTA, pure-tone average; Right, right-ear; WE, worse-ear.*

AE method which uses the threshold measures averaged over the two ears as the outcome, and the both-ear method, which uses the threshold measures from both of the two ears, may be preferred (Sheng et al. 2022). See Table 1 for an example of AE and both-ear outcomes. Note that, for presentational simplicity and to distinguish from the both-ear method which uses both ears' data as a cluster, we categorize AE as a single-ear method throughout this article even though its outcome is derived from both ears' data. In the presence of ear-level confounders, failure to include them in the model would lead to biased estimates for all methods. Previous studies demonstrate that if the ear-level confounders are included in the model, then the WE and BE methods may still lead to biased estimators. However, in this case, the bias would be less than that in models where the ear-level confounders were not included. Moreover, the AE method may lead to unbiased, but less efficient, estimators compared with the both-ear method (Sheng et al. 2022). On the other hand, biased estimators do not necessarily imply invalid hypothesis tests (Stahlecker & Schmidt 1996). A valid test means that the type I error rate, the probability of rejecting the null hypothesis when the null hypothesis is true, is controlled under the significance level, which is typically set at 5%. A test is invalid if the type I error rate is greater than the significance level. Similarly, estimators with higher efficiency do not necessarily imply hypothesis tests with higher power (Sundrum 1954). A powerful test means that the type II error rate, the probability of not rejecting the null hypothesis when the alternative hypothesis is true, is small. More specifically in terms of mathematical formula, statistical power = (1 – type II error). In commonly used hearing threshold data analysis methods, PTA measurements based on different groups of frequencies (e.g., low-frequency PTA, mid-frequency PTA, high-frequency PTA) are analyzed separately in different models instead of simultaneously in the same model. Thus, if one concludes that there is an exposure-outcome association because at least one of the tests in the separate analyses for individual frequency groups is significant, then issues with multiple comparisons and the correlations between the estimated parameters for these frequency groups are ignored, leading to an inflated type I error. In this case, one can use adjustment strategies to account for multiple comparisons, such as the Bonferroni method. However, these adjustment methods may not be able to easily address highly correlated hypotheses (Chen et al. 2017).

Previous studies that compared the single-ear and both-ear methods focused only on estimators for the exposure-hearing association (Chen et al. 2022; Sheng et al. 2022), while less concerned about hypothesis testing. In this study, our objective is to evaluate the validity and power of the statistical tests to evaluate exposure-hearing associations using both-ear and single-ear methods. We consider two methods for hypothesis tests based on both-ear data: (1) methods based on regression analyses inclusive of all frequencies in the same model, and (2) methods based on separate regression analyses where each model evaluates a single frequency. For the single-ear methods, we consider WE, BE, left/right-ear method, and AE; all frequencies are included in the same model for all the single-ear methods. We will consider both hypothesis tests for each frequency and a global test across all frequencies. The generalized estimating equation (GEE) method is used to obtain estimates and their variances for the regression coefficients of the regression models, taking into account the within-participant correlation between hearing threshold measurements in both ears and across frequencies, if applicable.

## STATISTICAL MODELS AND METHODS

When only one frequency group (i.e., low-frequency PTA, mid-frequency PTA, high-frequency PTA) is included in the analysis, we can assume the following both-ear model:

$$E\left(Y_{i,j}|X_i,\ W_i,\ Z_{i,j}\right) = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 Z_{i,j}$$

where $i$ indexes participants ($i = 1, 2, 3, \ldots, N$), $j$ is the index for the ear ($j = 0, 1$), and $Y_{i,j}$ is the hearing threshold measurement for individual $i$ and ear $j$. Among the independent variables, $X$ represents the participant-level exposure of interest, $W$ is possibly vector-valued participant-level potential confounders, and $Z$ represents possibly vector-valued ear-specific potential confounders (e.g., baseline hearing thresholds). Without further specification, parameters are row vectors and variables are column vectors throughout this article.

If multiple frequencies are considered, we need to add covariate-frequency interactions in the both-ear model, leading to

$$E\left(Y_{i,j,q}|X_i,\ W_i,\ Z_{i,j}\right) = \beta_{0q} + \beta_{1q} X_i + \beta_{2q} W_i + \beta_{3q} Z_{i,j} \quad (1)$$

where $E(\cdot)$ stands for the expected value, $q$ $(q = 1, \ldots Q)$ indexes the pure-tone frequency, $\beta_{1q}$ represents the association of the exposure with the hearing threshold level under the $q$ th frequency category, and $\beta_1$, the collection of $\beta_{1q}$, $q = 1, \ldots Q$, is a vector of parameters of interest.

For the single-ear method, we consider the WE, BE, left/right-ear method and AE. The models for single-ear method are similar to model (Eq. 1), deleting the subscript for ear, $j$, and replacing outcomes and covariates from both ears with those from a single-ear or an average from two ears. Using WE as an example, the outcome variable, $Y_{i,j,q}$, in model (Eq. 1) is now the WE outcome, $Y_{i,q}$, the participant-level covariates $X_i$ and $W_i$ stay the same, and the ear-level covariate $Z_{i,j}$ is now $Z_i$, taking the value for the WE. For AE, the outcome is now the AE outcome

and the ear-level covariate is now the average of the covariates between the two ears. Detailed models with mathematical notations are presented in Appendix A in Supplemental Digital Content 1, http://links.lww.com/EANDH/B373.

To evaluate whether there is an exposure-hearing threshold association, a commonly used method is to fit models for each frequency separately and test for $H_0 : \beta_{1q} = 0$ based on the model for the $q$th frequency, $q = 1, \ldots, Q$; if one of these tests is rejected, it is concluded that there is an exposure-hearing threshold association for at least one frequency. As shown in Appendices A and B in Supplemental Digital Content 1, http://links.lww.com/EANDH/B373, we have the following conclusions for the Wald test based on the single-ear method for testing $H_0 : \beta_{1q} = 0$ for a given $q$:

1. For the tests based on WE or BE, if there are no ear-level confounders, they are valid but less powerful than the both-ear method; however, if there are ear-level confounders, they may be invalid.
2. The tests based on the left/right-ear method, either in the presence or absence of ear-level covariates, are valid but less powerful compared with the both-ear method.
3. The tests based on AE are valid either in the presence or absence of ear-level covariates. They are as efficient as the both-ear method in the absence of ear-level confounders; however, in the presence of ear-level confounders, AE may lead to less powerful hypothesis tests than the both-ear method.

Note that the approach of testing for $H_0 : \beta_{1q} = 0$ for each $q$ and concluding that there is an exposure-hearing threshold association for at least one frequency if one of these tests is rejected may lead to an inflated type I error rate due to multiple comparisons. One of the correct methods is to test for the null hypothesis $H_{0g} : \beta_{11} = \ldots = \beta_{1Q} = 0$ based on model (Eq. 1), where the subscript $g$ denotes global tests; the alternative hypothesis is $H_{1g} : \beta_{1q} \neq 0$ for at least one frequency. This represents a global test in that it controls an overall type I error rate across all the frequencies. If this global test is significant, we can then further evaluate for which specific frequency the association is significant. To obtain the test statistic for $H_{0g}$, we fit model (Eq. 1) including all the frequencies, and estimate the variance-covariance matrix of the estimated regression coefficients that represent the associations of the exposure with the hearing threshold outcome at each of the individual frequencies. By contrast, fitting models for each frequency separately and assuming the covariance between $\hat{\beta}_{1q}$ being 0 may result in an inflated type I error rate.

Because threshold measurements between ears, between frequencies, and between time points, if applicable, are correlated, we choose to use the GEE approach (Zeger & Liang 1986) to make inferences about the regression parameters, $\beta = (\beta_{0q}, \beta_{1q}, \beta_{2q}, \beta_{3q},\ q = 1, \ldots Q)$, in model (Eq. 1). Similar to the (generalized) linear mixed effects model (GLMM/LMM) and multivariate analysis of variance methods, GEE can handle correlated data. Multivariate analysis of variance is appropriate only for categorical independent variables, whereas GEE and GLMM/LMM do not have restrictions on variable types. One advantage of GEE over LMM is that it requires fewer assumptions regarding the distribution of the data. For example, LMM assumes that random effects follow a normal distribution with mean zero (Gardiner et al. 2009), and accuracy of the inference (i.e., SE) relies on correct specification of both the underlying mean model and error distribution, while GEE only requires the mean model being correctly specified for a valid inference (Hubbard et al. 2010). We can specify an exchangeable working variance-covariate matrix for GEE, to take the between-ear correlations into account. Twisk (2004) pointed out that, for continuous outcomes, GEE with an exchangeable working correlation structure often gives comparable results to a random intercept linear model, when both models are correctly specified. By the standard asymptotic theory of GEE, the $\beta$-estimator, $\hat{\beta}$, converges to the true value as the sample size increases, as long as the outcome mean is modeled correctly. See Appendix C in Supplemental Digital Content 1, http://links.lww.com/EANDH/B373, for more technical details of GEE.

## SIMULATION STUDY

Monte Carlo simulation studies are often useful for assessing the performance of statistical methods, where pseudo-data are generated so that we know and can control the true models and model parameters that are used to generate the data. We can then compare the analysis results of the generated data with the truth, based on which the data are generated. Here, we conducted a simulation study to compare the both-ear (based on model [Eq. 1]), left-ear, WE, BE, AE methods, where all frequency groups were included in each model, as well as the both-ear method that fitted models for each frequency group separately, which for brevity is referred to as the both-ear separate method. We simulated 1000 samples each composed of 1000 subjects' left- and right-ear PTA threshold measurements at low- and high-frequency groups. The outcome data were generated based on model $E(Y) = \beta_1 x(1-q) + \beta_2 xq + 0.3z(1-q) + 0.25zq + 0.2q$, where $x$ was the participant-level exposure, generated from $N(3,\ 4)$, a normal distribution where the mean is 3 and the SD is 4, $q$ was the binary frequency indicator ($q$ = 0 for low and 1 for high), $z$ was an ear-level covariate, assuming mean zero, and $\beta_1$ and $\beta_2$ represented the $x - Y$ association for low and high frequencies, respectively. On the basis of the preliminary results from the analysis of Conservation of Hearing Study (CHEARS)-AAA study data, the outcome data were generated either under the null $\beta_1 = \beta_2 = 0$, or alternatively assuming $\beta_1 = 0.08$ and $\beta_2 = 0.1$ following a multivariate normal distribution, with the following correlations: $\rho = 0.7$ for the outcomes between two ears at the same frequency; $\rho = 0.4$ for the outcomes for the same ear between different frequencies, and $\rho = 0.3$ for the outcomes between two ears at different frequencies. We have rounded the simulated outcome data to five units (e.g., 5, 15, 20, 25 dB) to mimic clinical hearing threshold measurements, which are obtained in 5 dB increments.

We consider two scenarios. In scenario I, the ear-level covariate, $z$, was independent of the exposure, and thus it was not a confounder. In scenario II, $z$ was an ear-specific confounder, with the correlation coefficient between $z$ and $x$ equal to 0.2.

We used GEE to estimate the regression coefficients and performed the following tests. (1) Test 1: Wald test for the null hypothesis $H_{01} : \beta_1 = 0$; that is, assessing whether the exposure is associated with low-frequency (i.e., $q$ = 0) thresholds; (2) test 2: Wald test for $H_{02} : \beta_2 = 0$; that is, assessing whether the exposure is associated with high frequency (i.e., $q$ = 1) thresholds; (3) test 3: if either test 1 or test 2 is rejected, a conclusion is drawn that there was an exposure-hearing

threshold association for at least one frequency. Note that this approach is commonly used in practice and may lead to an inflated type I error rate. The recommended method for this test is (4) test 4 (a global test): Wald test for the null hypothesis $H_{0g}: \beta_1 = \beta_2 = 0$; that is, assessing whether there is an exposure-hearing threshold association for at least one frequency group. In addition, we fitted random intercept LMMs to better illustrate the difference between GEE and LMM. For the tests listed earlier we evaluated the type I error rates, which are the percentages of rejecting the null hypothesis among the simulation replicates when the data were generated assuming there was no exposure-hearing associations, and powers, the percentages of rejecting the null hypothesis when the data were generated under the alternative hypotheses. These two metrics are essential for ensuring the validity and efficiency of the tests. The test is valid only if its type I error rate is less than 5%. If the power of the test is low, it will be less likely to detect a true effect even if it indeed exists. In addition, to validate the results of previous studies (Sheng et al. 2022), we compared the relative bias, defined as $(\hat{\beta}_q - \beta_q)/\beta_q,\ q = 1, 2$, or bias if the true coefficient is zero, defined as $\hat{\beta}_q - \beta_q$, the empirical SD (Knudsen et al. 2016) of the point estimates over the simulation replicates, and 95% confidence interval coverage rates between the methods. Smaller empirical SD typically implies more efficient estimates, and a coverage rate closer to 95% indicates better interval estimates.

The simulation results for scenario I are shown in Table 2, which presents the type I error rates and powers for tests, and Supplementary Tables 1 and 2 in Supplemental Digital Content 2, http://links.lww.com/EANDH/B374, which shows the estimates. As expected, the both-ear method (joint), and AE outperformed WE, BE, left-ear, and both-ear (separate) method. Test 3 based on any of the methods had an inflated type I error rate. The both-ear method with separate analysis for each individual frequency also had an inflated type 1 error rate for the global test (test 4). The other tests all had type I error rates of approximately the significance level 5%. The statistical power of each single-ear method (WE, BE, AE, left-ear method) was lower, ranging from 3 to 24% lower, than the power of both ear and AE. Regarding the point estimates in Supplementary Tables 1 and 2 in Supplemental Digital Content 2, http://links.lww.com/EANDH/B374, the both-ear estimator was more efficient than that of WE/BE/left-ear method. AE had similar performance to the both-ear method for both tests and estimates. This is consistent with previous findings of Sheng et al. (2022).

Table 3, which presents the type I error rates and powers for tests, and Supplementary Tables 3 and 4 in Supplemental Digital Content 2, http://links.lww.com/EANDH/B374, which present the estimates, show the simulation results for scenario II, where there was an ear-specific confounder. The both-ear method outperformed WE, BE, AE, and the left-ear method. Tests 1 and 2 using WE and BE, test 3 using all methods, and test 4 using WE and BE and the both-ear method for each individual frequency separately all had inflated type I error rates. For low-frequency, the relative bias for WE and BE was larger than 30%; for high frequency, the relative bias was greater

**TABLE 2. Comparison of type I error rates and powers of tests between different methods for true model $E(Y) = \beta_1 x(1-q) + \beta_2 xq + 0.3z(1-q) + 0.25zq + 0.2q$ under significance level 5% (scenario I)**

| | Both Ear, Joint | Left Ear, Joint | WE, Joint | BE, Joint | AE, Joint | Both Ear, Separate |
|---|---|---|---|---|---|---|
| Type 1 error rate ($\beta_1 = \beta_2 = 0$) | | | | | | |
| Test 1. $H_{01}$: $\beta_1$ =0 | 5.3% | 5.2% | 5.7% | 5.4% | 4.4% | 4.6% |
| Test 2. $H_{02}$: $\beta_2$ =0 | 4.8% | 5.3% | 4.9% | 5.4% | 4.8% | 4.6% |
| Test 3. At least one of tests 1, 2 rejected | 8.6% | 8.8% | 10.3% | 10.8% | 8.5% | 8.4% |
| Test 4. $H_{0g}$: $\beta_1 = \beta_2$ =0 | 5.6% | 5.4% | 5.9% | 5.7% | 4.8% | 7.2% |
| Power ($\beta_1 = 0.08; \beta_2 = 0.1$) | | | | | | |
| Test 1. $H_{01}$: $\beta_1$ =0 | 84.2% | 69.4% | 64.1% | 76.6% | 82.8% | 84.2% |
| Test 2. $H_{02}$: $\beta_2$ =0 | 95.6% | 87.5% | 86.0% | 71.0% | 95.0% | 97.1% |
| Test 4. $H_{0g}$: $\beta_1 = \beta_2$ =0 | 92.6% | 79.8% | 89.3% | 88.9% | 92.0% | — |

*The ear-level covariate, z, is independent of the exposure, and therefore it is not a confounder.*
*AE, average-ear; BE, better-ear; Joint, the model includes both low- and high-frequency groups and therefore $\beta_1$ and $\beta_2$ are estimated in the same model, $E(Y) = \beta_1 x(1-q) + \beta_2 xq + \beta_3 z(1-q) + \beta_4 zq + \beta_5 q$; Separate, the model is fitted for each frequency group separately, $E(Y) = \beta_1 x + \beta_3 z$, and $E(Y) = \beta_2 xq + \beta_4 zq + \beta_5 q$; WE, worse-ear.*

**TABLE 3. Comparison of type I error rates and powers of tests between different methods for model $E(Y) = \beta_1 x(1-q) + \beta_2 xq + 0.3z(1-q) + 0.25zq + 0.2q$ under significance level 5% (scenario II)**

| | Both Ear, Joint | Left-Ear, Joint | WE, Joint | BE, Joint | AE, Joint | Both Ear, Separate |
|---|---|---|---|---|---|---|
| Type 1 error rate ($\beta_1 = \beta_2 = 0$) | | | | | | |
| Test 1. $H_0$: $\beta_1$ =0 | 3.9% | 4.5% | 8.0% | 7.2% | 3.3% | 4.2% |
| Test 2. $H_0$: $\beta_2$ =0 | 4.7% | 4.3% | 9.3% | 9.5% | 5.2% | 5.4% |
| Test 3. At least one of tests 1, 2 rejected | 7.3% | 7.1% | 16.4% | 15.8% | 7.4% | 8.1% |
| Test 4. $H_0$: $\beta_1 = \beta_2$ =0 | 4.4% | 5.5% | 9.0% | 8.8% | 4.4% | 7.5% |
| Power ($\beta_1 = 0.08; \beta_2 = 0.1$) | | | | | | |
| Test 1. $H_0$: $\beta_1$ =0 | 45.6% | 33.3% | — | — | 43.5% | 45.7% |
| Test 2. $H_0$: $\beta_2$ =0 | 74.5% | 57.5% | — | — | 74.5% | 73.3% |
| Test 4. $H_0$: $\beta_1 = \beta_2$ =0 | 63.2% | 46.4% | — | — | 63.3% | — |

*The ear-level covariate, z, is independent of the exposure, and therefore it is not a confounder.*
*AE, average-ear; BE, better-ear; Joint, the model includes both low- and high-frequency groups and therefore $\beta_1$ and $\beta_2$ are estimated in the same model, $E(Y) = \beta_1 x(1-q) + \beta_2 xq + \beta_3 z(1-q) + \beta_4 zq + \beta_5 q$; Separate, the model is fitted for each frequency group separately, $E(Y) = \beta_1 x + \beta_3 z$, and $E(Y) = \beta_2 xq + \beta_4 zq + \beta_5 q$; WE, worse-ear.*

than 15%. In addition, the left-ear methods resulted in lower statistical power and higher variance. We did not consider the power of WE and BE, as the tests using these methods were invalid. When evaluating the association of the exposure with low-frequency thresholds, the statistical power using the left-ear method was 15% lower than the both-ear method. Although slightly less efficient, the results obtained using AE were comparable to the both-ear method.

Note that the powers of hypothesis tests depend on value $\beta$, which represents the level of exposure-hearing associations in the data, in addition to sample size, cluster size, etc. For example, when the number of observations is fixed, a larger cluster size or a weaker exposure-hearing association typically leads to lower power. Therefore, the powers displayed in Tables 2 and 3 may not be generalized to a different scenario.

Results of the random intercept LMMs fitted for scenario I and scenario II are shown in Supplementary Tables 5 and 6 in Supplemental Digital Content 2, http://links.lww.com/EANDH/B374. In scenario I, the type I error rates and power are similar between the models fitted using GEE and random intercept LMMs for all of the methods. This is consistent with the previous finding that the results of GEE with exchangeable working correlation matrix is similar to those of random intercept LMMs when both models are correctly specified (Twisk 2004). In scenario I, where the ear-level covariate is correlated with participant-level covariates, each method fitted using an LMM with random intercepts has roughly 20% less power compared with the corresponding methods fitted by GEE. This loss of power might be due to the fact the correlation structure used in the data generation does not align with the model assumptions of the random intercept LMM.

We also examine the effect of varying sample sizes on type I error and power of the test using estimates from GEE in scenario II, and present simulation results in Supplementary Tables 7 and 8 in Supplemental Digital Content 2, http://links.lww.com/EANDH/B374. The sample size of Supplementary Tables 7 and 8 in Supplemental Digital Content 2, http://links.lww.com/EANDH/B374, is 200 and 4000, respectively. When sample size equal to 200, type I error rates of all the methods are higher than the expected level (i.e., 5%), ranging from 7.8 to 12.4% in tests 1, 2, and 4. In addition, power of all tests are very low, roughly at 15%. On the other hand, type I error rates of all methods when the sample size is 4000 are generally controlled under 5% in tests 1, 2, and 4. Power of the tests is much higher than that when sample size is 200 as expected.

## CHEARS DATA ANALYSIS

Next, we illustrate the difference between these methods using a real-world data example. The CHEARS is an epidemiological study that investigates relations between a number of medical, dietary, and other lifestyle factors and the risk of hearing loss among several large, well-characterized longitudinal cohorts, including the Nurses' Health Study II (Curhan et al. 2013). A subcohort of Nurses' Health Study II participants was invited to undergo comprehensive CHEARS clinical audiometry assessments conducted by licensed audiologists according to a rigorous set of standardized protocols and procedures (Curhan et al. 2019). In total, 3749 participants completed baseline audiometry and 3135 participants (84%) completed 3-year follow-up testing. Detailed information on CHEARS has been described previously (Curhan et al. 2019, 2021). In this example, the outcome was a continuous variable, defined as the elevation in PTA hearing threshold measurements from baseline to year 3. For each participant, the outcome variable had six dimensions, PTA at low-frequency (0.5, 1, 2 kHz), mid-frequency (3, 4 kHz), and high-frequency (6, 8 kHz) for each individual ear. The exposure of interest was the consumption of a healthful diet, as assessed by the Dietary Approaches to Stop Hypertension (DASH) dietary adherence score. A higher DASH score indicated greater adherence to the recommendations provided for the DASH diet (i.e., "healthier diet"), and DASH scores were categorized in quartiles according to the distribution in the study population.

For illustration purposes, we adjusted for baseline PTA measurements and the following potential confounders: age (continuous), race (black/white/multi/other or unknown), body mass index (<25 kg/m$^2$/25 to 29/30 to 34/35 to 39/40+), smoking status (never/past/current), history of tinnitus (yes/no), cumulative average energy intake (continuous), noise exposure (very loud occupational or leisure-time noise exposure ≥3 hours/week during any decade; yes/no). We also included interaction terms between the DASH score and three frequency levels: low-frequency ($PTA_{0.5,1,2\ kHz}$), mid-frequency ($PTA_{3,4\ kHz}$), and high-frequency ($PTA_{6,8\ kHz}$). GEE with an exchangeable covariance matrix was used in all the methods.

Table 4 shows the results for the multivariable-adjusted mean difference in hearing threshold elevation between each of the second to fourth quartiles of the DASH score and the first quartile (the reference category). These results are shown in hearing threshold elevation at low-, mid-, and high-frequencies, corresponding to the three PTA outcomes, and they were calculated based on the regression coefficient estimates of the DASH score and interactions between the DASH score and frequency. The $p$ values for the tests were for linear trend, where the median DASH score in each quartile was treated as a continuous variable. This linear trend test tested whether the association between the DASH score and hearing threshold elevation increased linearly with higher quartiles of DASH score. The global test for evaluating whether there was a linear association of DASH score with hearing threshold elevation at any of the low-, mid-, and high-frequency PTA outcomes was significant or marginally significant when using the both-ear ($p = 0.04$) and AE method ($p = 0.05$), but not when using the left- and right-ear methods. Frequency-specific tests from the both-ear method show that the linear association of DASH score with the hearing threshold elevation might lie at low- ($p = 0.041$) and mid-frequencies ($p = 0.005$).

Supplementary Table 9 in Supplemental Digital Content 2, http://links.lww.com/EANDH/B374, presents the results of the regression models above fitted by random intercept LMMs instead of GEE. In terms of both-ear (joint) method, the coefficient estimates from these two models are close. The $p$ values for trend test given by GEE are lower than 5% in all frequency groups while the $p$ values obtained by random intercept LMMs are lower than 5% only in the medium frequency group. With regards to the single-ear methods, the coefficient estimates from random intercept LMMs and GEE are mostly the same, while the SE estimates are larger in LMMs. Therefore, the trend tests using GEE generally have more significant $p$ values than that using random intercept LMMs.

**TABLE 4. The mean difference (in dB) between DASH quartiles in 3-yr audiometric hearing threshold elevation in the CHEARS audiology assessment arm**

| | Both Ear, Joint | | Left-Ear, Joint | | Right-Ear, Joint | | WE, Joint | | BE, Joint | | AE, Joint | | Both Ear, Separate | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | 95%CI | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI |
| Low-frequency | | | | | | | | | | | | | | |
| Q1 | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference |
| Q2 | −0.02 | (−0.37 to 0.32) | 0.21 | (−0.20 to 0.61) | −0.23 | (−0.63 to 0.16) | 0.17 | (−0.22 to 0.56) | −0.17 | (−0.53 to 0.19) | 0.00 | (−0.34 to 0.34) | −0.01 | (−0.36 to 0.33) |
| Q3 | −0.28 | (−0.62 to 0.05) | −0.16 | (−0.55 to 0.23) | −0.37 | (−0.76 to 0.02) | −0.18 | (−0.54 to 0.19) | −0.31 | (−0.67 to 0.04) | −0.24 | (−0.57 to 0.09) | −0.24 | (−0.58 to 0.10) |
| Q4 | −0.33 | (−0.70 to 0.04) | −0.28 | (−0.69 to 0.13) | −0.34 | (−0.78 to 0.09) | −0.18 | (−0.58 to 0.22) | −0.39 | (−0.76 to −0.01) | −0.27 | (−0.63 to 0.09) | −0.29 | (−0.66 to 0.07) |
| Trend test *p* value | | 0.041 | | 0.090 | | 0.099 | | 0.198 | | 0.039 | | 0.080 | | 0.066 |
| Medium frequency | | | | | | | | | | | | | | |
| Q1 | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference |
| Q2 | 0.08 | (−0.38 to 0.55) | 0.23 | (−0.31 to 0.76) | −0.06 | (−0.61 to 0.49) | 0.20 | (−0.33 to 0.73) | −0.04 | (−0.51 to 0.43) | −0.07 | (−0.38 to 0.53) | 0.12 | (−0.34 to 0.58) |
| Q3 | −0.17 | (−0.63 to 0.29) | −0.06 | (−0.60 to 0.47) | −0.27 | (−0.81 to 0.27) | −0.12 | (−0.63 to 0.38) | −0.20 | (−0.69 to 0.28) | −0.16 | (−0.62 to 0.29) | −0.11 | (−0.58 to 0.35) |
| Q4 | −0.65 | (−1.12 to −0.17) | −0.62 | (−1.15 to −0.09) | −0.67 | (−1.25 to −0.09) | −0.61 | (−1.14 to −0.08) | −0.68 | (−1.16 to −0.20) | −0.64 | (−1.10 to −0.18) | −0.56 | (−1.04 to −0.07) |
| Trend test *p* value | | 0.005 | | 0.016 | | 0.018 | | 0.015 | | 0.006 | | 0.005 | | 0.018 |
| High-frequency | | | | | | | | | | | | | | |
| Q1 | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference | 0 | Reference |
| Q2 | −0.20 | (−0.81 to 0.41) | −0.08 | (−0.83 to 0.67) | −0.30 | (−1.02 to 0.41) | −0.27 | (−0.99 to 0.45) | −0.11 | (−0.72 to 0.50) | −0.17 | (−0.78 to 0.43) | −0.24 | (−0.85 to 0.38) |
| Q3 | −0.21 | (−0.84 to 0.43) | −0.28 | (−1.04 to 0.49) | −0.10 | (−0.83 to 0.62) | −0.29 | (−1.01 to 0.43) | −0.07 | (−0.71 to 0.58) | −0.16 | (−0.79 to 0.46) | −0.30 | (−0.96 to 0.35) |
| Q4 | −0.46 | (−1.11 to 0.19) | −0.72 | (−1.50 to 0.07) | −0.15 | (−0.92 to 0.62) | −0.30 | (−1.06 to 0.45) | −0.58 | (−1.18 to 0.14) | −0.39 | (−1.03 to 0.25) | −0.61 | (−1.30 to 0.09) |
| Trend test *p* value | | 0.187 | | 0.070 | | 0.817 | | 0.442 | | 0.158 | | 0.262 | | 0.096 |
| Global test *p* value | | 0.041 | | 0.074 | | 0.096 | | 0.111 | | 0.040 | | 0.045 | | Invalid |

*AE, average-ear; BE, better-ear; CHEARS, Conservation of Hearing Study; CI, confidence interval; DASH, Dietary Approaches to Stop Hypertension; Joint, the model includes both low- and high-frequency groups and therefore $\beta_1$ and $\beta_2$ are measured in the same model; Separate, the model is fitted for each frequency group separately, and therefore $\beta_1$ and $\beta_2$ are measured in the different model separately; WE, worse-ear.*

## DISCUSSION

In this article, we demonstrate that using the both-ear GEE method that includes threshold measurements at all audiometric frequencies in the same model may be the preferred strategy for testing whether there is an exposure-hearing threshold association (the global test). This method offers several distinctive advantages that address the limitations of commonly used methods in previous literature. Specifically, (1) in the absence of ear-level confounders, WE, BE, and left/right-ear method have less power for frequency-specific tests and for the global test; (2) in the presence of ear-level confounders, WE and BE are invalid, and the left/right-ear and AE methods have less power, with the power loss in the left/right-ear methods much greater than the AE method, for frequency-specific tests and for the global test; and (3) the both-ear method with separate analyses for each individual frequency is invalid for the global test.

Note that in the illustrative example based on the CHEARS-AAA data, a potential ear-level confounder, baseline PTA measurement, is included in the model. Therefore, the tests based on WE and BE may be invalid. On the basis of the results presented in Table 4, the global test using both-ear method and AE method showed a significant exposure-outcome association for at least one of the PTA frequency groups while left/right-ear method failed to do so. This suggests that there may be a true association between DASH score and hearing threshold elevation and that the left/right-ear methods were not powerful enough to detect the association. In addition, the both-ear method typically provides smaller *p* values than the AE method when conducting the trend tests or the global test. This is consistent with our conclusion that the both-ear method typically has a better power than the AE method when there are ear-level covariates in the model. Hence, when analyzing hearing threshold data, it is preferable to use the both-ear method that includes all of the tested audiometric frequencies in the model, to have a valid and a more powerful test, especially when ear-level confounders are present. The AE method is also valid but may be less powerful. On the other hand, the WE and BE methods should not be used because they may fail to control the type I error rate to below 5%. The methods and conclusions would be the same for analyses using an outcome definition based on individual threshold measurements as they would be for analyses using an outcome definition based on PTA measurements.

While previous papers (Chen et al. 2022; Sheng et al. 2022) focus on estimators from WE, BE, both-ear, and AE methods, our article further compares type I error and statistical power of different hypothesis tests based on these methods. Our findings are consistent with those from Sheng et al. (2022) and demonstrate that with or without ear-level confounders, the both-ear method that includes all frequencies in the same model is recommended for both estimating the exposure-outcome association and hypothesis testing.

The GEE method requires specification of the structure of the correlation matrix for the clustered outcomes. A misspecified working correlation matrix still leads to valid regression coefficient estimates. However, if the correlation structure is correctly specified, it will typically increase the efficiency of the estimator (Fitzmaurice 1995), and the power of the test. We used an exchangeable working correlation structure in this article to account for the within-participant between-ear and between frequency correlations. An unstructured working correlation structure can be used in sensitivity analyses.

In addition, there are some limitations of GEE methods when the data structure is complex, such as its inability to accurately model data from multiple non-nested sources of clustering (Betensky et al. 2000).

We also fitted a random intercept LMM to compare with the results given by GEE. When there is no ear-level confounder, these two models generally produce very similar type I error rates and power. However, with the presence of ear-level confounder, GEE is more powerful than random intercept LMMs, probably because the assumptions of the latter are violated, such as correct specification of the mean model and distribution of the errors (Hubbard et al. 2010).

Wald tests rely on the asymptotic distribution theory (i.e., large sample size). Thus, as shown in our simulation study, a small sample size may lead to inflated type I error rates. In this case, it is recommended to perform finite-sample adjusts to Wald-type tests to constrain type I error rate under the expected level (Fay & Graubard 2001).

Notably, certain risk factors may be associated with changes in hearing thresholds in only one ear, such as a unilateral ear injury or unilateral chronic otitis media. Nevertheless, in scenarios where an ear-level risk factor may be differentially related to changes in hearing thresholds in the individual ears, we can include an exposure-ear interaction term in the models and the both-ear method would still be the preferred analytic approach. The findings in this article also apply to those scenarios.

## CONCLUSION

For hypothesis testing to evaluate whether there are significant associations between an exposure of interest and audiometric hearing threshold measurements, the both-ear method that includes all frequencies in the same model is the recommended analytic approach.

## ACKNOWLEDGMENTS

This work was partially supported by the National Institute Health grants R01 DC017717, U01 CA176726 (NHS II), and U01 DC 010811.

Z.L. conducted the simulation study and wrote the first draft of the manuscript; J.L. performed the data analysis; M.W. developed the analytical methods; S.C and G.C. contributed to the study conception and design and data acquisition; Z.W. conducted additional simulation studies and real data analysis for the revisions; M.W., Z.W., and S.C. provided critical revision of the manuscript. Z.L. and Z.W. are the co-first authors. S.C. and M.W. are the co-senior authors. All authors approved the final version and made the decision to submit the manuscript for publication.

G. C. receives royalties from UpToDate for being an author and Section Editor; and is an employee of OM1, Inc. The other authors have no disclosures.

Address for correspondence: Molin Wang, Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, 665 Huntington Ave, Boston, MA 02115, USA. E-mail: mwang@hsph.harvard.edu

Received January 25, 2023; accepted February 23, 2024; published online ahead of print March 28, 2024

## REFERENCES

Bainbridge, K. E., Hoffman, H. J., Cowie, C. C. (2008). Diabetes and hearing impairment in the United States: Audiometric evidence from the National Health and Nutrition Examination Survey, 1999 to 2004. *Ann Intern Med*, *149*, 1–10.

Betensky, R. A., Talcott, J. A., Weeks, J. C. (2000). Binary data with two, non-nested sources of clustering: An analysis of physician recommendations for early prostate cancer treatment. *Biostatistics*, *1*, 219–230.

Chadha, S., Kamenov, K., Cieza, A. (2021). The world report on hearing, 2021. *Bull World Health Organ*, *99*, 242–242A.

Chen, S. Y., Feng, Z., Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *J Thorac Dis*, *9*, 1725–1729.

Chen, C., Zhang, N., Curhan, G. C., Curhan, S. G., Wang, M. (2022). Both-ear method for the analysis of audiometric data. *Ear Hear*, *43*, 1447–1455.

Cruickshanks, K. J., Tweed, T. S., Wiley, T. L., Klein, B. E., Klein, R., Chappell, R., Nondahl, D. M., Dalton, D. S. (2003). The 5-year incidence and progression of hearing loss: The epidemiology of hearing loss study. *Arch Otolaryngol Head Neck Surg*, *129*, 1041–1046.

Curhan, G. C., Eavey, R., Wang, M., Stampfer, M. J., Curhan, G. C. (2013). Body mass index, waist circumference, physical activity, and risk of hearing loss in women. *Am J Med*, *126*, 1142.e1141–1142.e1148.

Curhan, G. C., Halpin, C., Wang, M., Eavey, R. D., Curhan, G. C. (2019). Prospective study of dietary patterns and hearing threshold elevation. *Am J Epidemiol*, *189*, 204–214.

Curhan, G. C., Halpin, C., Wang, M., Eavey, R. D., Curhan, G. C. (2021). Tinnitus and 3-year change in audiometric hearing thresholds. *Ear Hear*, *42*, 886–895.

Dawes, P., Emsley, R., Cruickshanks, K. J., Moore, D. R., Fortnum, H., Edmondson-Jones, M., McCormack, A., Munro, K. J. (2015). Hearing loss and cognition: The role of hearing AIDS, social isolation and depression. *PLoS One*, *10*, e0119616.

Deng, A. (2012). Hypothesis testing when a linear regression is estimated by biased estimators. *SSRN Electron J*.

Fay, M. P., & Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, *57*, 1198–1206.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, *51*, 309–317.

Gardiner, J. C., Luo, Z., Roman, L. A. (2009). Fixed effects, random effects and GEE: What are the differences? *Stat Med*, *28*, 221–239.

Grondin, Y., Bortoni, M. E., Sepulveda, R., Ghelfi, E., Bartos, A., Cotanche, D., Clifford, R. E., Rogers, R. A. (2015). Genetic polymorphisms associated with hearing threshold shift in subjects during first encounter with occupational impulse noise. *PLoS One*, *10*, e0130827.

Haile, L. M., Kamenov, K., Briant, P. S., Orji, A. U., Steinmetz, J. D., Abdoli, A., Abdollahi, M., Abu-Gharbieh, E., Afshin, A., Ahmed, H., Ahmed Rashid, T., Akalu, Y., Alahdab, F., Alanezi, F. M., Alanzi, T. M., Al Hamad, H., Ali, L., Alipour, V., Al-Raddadi, R. M., Chadha, S. (2021). Hearing loss prevalence and years lived with disability, 1990–2019: Findings from the Global Burden of Disease Study 2019. *Lancet*, *397*, 996–1009.

Hu, H., Sasaki, N., Ogasawara, T., Nagahama, S., Akter, S., Kuwahara, K., Kochi, T., Eguchi, M., Kashino, I., Murakami, T., Shimizu, M., Uehara, A., Yamamoto, M., Nakagawa, T., Honda, T., Yamamoto, S., Hori, A., Nishiura, C., Okazaki, H., Dohi, S. (2018). Smoking, smoking cessation, and the risk of hearing loss: Japan Epidemiology Collaboration on Occupational Health Study. *Nicotine Tob Res*, *21*, 481–488.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., Bruckner, T., Satariano, W. A. (2010). To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, *21*, 467–474.

Knudsen, A. B., Zauber, A. G., Rutter, C. M., Naber, S. K., Doria-Rose, V. P., Pabiniak, C., Johanson, C., Fischer, S. E., Lansdorp-Vogelaar, I., Kuntz, K. M. (2016). Estimation of benefits, burden, and harms of colorectal cancer screening strategies: Modeling study for the US Preventive Services Task Force. *JAMA*, *315*, 2595–2609.

Lin, Y.-Y., Wu, L.-W., Kao, T.-W., Wu, C.-J., Yang, H.-F., Peng, T.-C., Lin, Y.-J., Chen, W.-L. (2016). Secondhand smoke is associated with hearing threshold shifts in obese adults. *Sci Rep*, *6*, 33071.

Mick, P., Kawachi, I., Lin, F. R. (2014). The association between hearing loss and social isolation in older adults. *Otolaryngol Head Neck Surg*, *150*, 378–384.

Sheng, Y., Yang, C., Curhan, S., Curhan, G., Wang, M. (2022). Analytical methods for correlated data arising from multicenter hearing studies. *Stat Med*, *41*, 5335–5348.

Shih, J. H., Li, I. H., Pan, K. T., Wang, C. H., Chen, H. C., Fann, L. Y., Tseng, J. H., Kao, L. T. (2020). Association between anemia and auditory threshold shifts in the US Population: National Health and Nutrition Examination Survey. *Int J Environ Res Public Health*, *17*, 3916.

Stahlecker, P., & Schmidt, K. (1996). Biased estimation and hypothesis testing in linear regression. *Acta Appl Math*, *43*, 145–151.

Sundrum, R. M. (1954). On the relation between estimating efficiency and the power of tests. *Biometrika*, *41*, 542.

Twisk, J. W. (2004). Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. *Eur J Epidemiol*, *19*, 769–776.

Verschuur, C. A., Dowell, A., Syddall, H. E., Ntani, G., Simmonds, S. J., Baylis, D., Gale, C. R., Walsh, B., Cooper, C., Lord, J. M., Sayer, A. A. (2012). Markers of inflammatory status are associated with hearing threshold in older people: Findings from the Hertfordshire Ageing Study. *Age Ageing*, *41*, 92–97.

WHO. (2021). *Deafness and hearing loss*. Retrieved October 1, 2022, from https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

Wilson, B. S., Tucci, D. L., Merson, M. H., O'Donoghue, G. M. (2017). Global hearing health care: New findings and perspectives. *Lancet*, *390*, 2503–2515.

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, *42*, 121–130.