Communications for Statistical Applications and Methods
2020, Vol. 27, No. 2, 225–239

https://doi.org/10.29220/CSAM.2020.27.2.225
Print ISSN 2287-7843 / Online ISSN 2383-4757

Ranking subjects based on paired compositional data with application to age-related hearing loss subtyping

Jin Hyun Nam[1,a], Aastha Khatiwada[1,a], Lois J. Matthews[b], Bradley A. Schulte[b,c], Judy R. Dubno[b,c], Dongjun Chung[2,a]

[a]Department of Public Health Sciences, Medical University of South Carolina, USA;
[b]Department of Otolaryngology-Head and Neck Surgery, Medical University of South Carolina, USA;
[c]Department of Pathology and Laboratory Medicine, Medical University of South Carolina, USA

Abstract

Analysis approaches for single compositional data are well established; however, effective analysis strategies for paired compositional data remain to be investigated. The current project was motivated by studies of age-related hearing loss (presbyacusis), where subjects are classified into four audiometric phenotypes that need to be ranked within these phenotypes based on their paired compositional data. We address this challenge by formulating this problem as a classification problem and integrating a penalized multinomial logistic regression model with compositional data analysis approaches. We utilize Elastic Net for a penalty function, while considering average, absolute difference, and perturbation operators for compositional data. We applied the proposed approach to the presbyacusis study of 532 subjects with probabilities that each ear of a subject belongs to each of four presbyacusis subtypes. We further investigated the ranking of presbyacusis subjects using the proposed approach based on previous literature. The data analysis results indicate that the proposed approach is effective for ranking subjects based on paired compositional data.

Keywords: presbyacusis, composition, penalized logistic regression, elastic net

1. Introduction

Age-related hearing loss (presbyacusis) is a major public health concern due to a growing aging population, high prevalence among older adults, associated significant communication difficulties and health-related problems (Lin *et al.*, 2011). The underlying biological mechanisms associated with presbyacusis are complex and remain unclear. One of the reasons for this challenge is complications in the phenotype definition of presbyacusis. Various studies showed that presbyacusis is not a single category of impairment, but rather consists of multiple sub-categories related to differences in pathophysiology (Dubno *et al.*, 2013; Vaden *et al.*, 2017).

The current study was motivated by ongoing genetic studies of a large number of older adults whose audiograms have been classified into one of four subtypes. In genetic studies, it is critical to have accurate phenotype definitions that differentiate subjects according to presumed underlying pathologies, in order to maximize statistical power, minimize biases, and demonstrate phenotype-genotype associations. In statistical genetics, extreme discordant phenotype (EDP) design has been

[1] Contributed equally to the work.
[2] Corresponding author: Department of Public Health Sciences, Medical University of South Carolina,135 Cannon Street Suite 303, Charleston, SC 29425, USA. E-mail: chungd@musc.edu

Published 31 March 2020 / journal homepage: http://csam.or.kr

used to address problems that arise when a wide range of classification accuracy within each phenotype is observed. In the EDP design, only subjects with purer phenotypes are used while subjects with ambiguous (impure) phenotypes are discarded. This approach has been reported to improve signals by "purifying" subjects (removing "noise" subjects) despite the loss of sample size (Barnett *et al.*, 2013; Huang and Lin, 2007; Li *et al.*, 2011; Zhang *et al.*, 2006).

Key variables of the presbyacusis dataset considered in this paper are the 8 variables indicating the probabilities that a subject's right and left ears belong to one of the four presbyacusis subtypes (Older-Normal, Metabolic, Sensory, and Metabolic + Sensory), generated using the supervised machine learning approach (Dubno *et al.*, 2013). Hence, we have a pair (both ears of the same subject) of compositional data (presbyacusis subtype probabilities) for each subject. Analysis approaches of compositional data have been investigated (Van den Boogaart and Tolosana-Delgado, 2013) with pioneering work by Aitchison (1982). These approaches include isometric log-ratio (ilr), centered log-ratio (clr) and additive log-ratio (alr) transformations. Therefore, popular statistical approaches can be used such as linear models, clustering analysis, and classification analysis (Van den Boogaart and Tolosana-Delgado, 2013) because the transformations remove the 'sum-to-one' constraint of compositional data and provide one-to-one mapping so that the model fitting results can be transformed back to original data space. Alternative approaches based on Dirichlet distributions were recently proposed to address issues of heteroscedasticity and challenges in interpretability (Maier, 2014). When we have a paired compositional data, we might use these transformation approaches for compositional data of each side, apply standard statistical approaches to each transformed compositional data, and then combine the analysis results between the pair. However, these approaches are still limited because compositional data is analyzed independently and does not take into account the pairing structure. To the authors' best knowledge, statistical strategies to analyze paired compositional data are currently lacking. This complication in identifying subjects with purer phenotypes introduces additional challenges when implementing the EDP design.

In this paper, we investigate the problem of purifying subjects by formulating it as a supervised learning problem, integrating the transformation of paired compositional data with the penalized multinomial logistic regression approaches, and by ranking subjects using predictive probabilities indicating the phenotype purity. We further evaluated the ranking of presbyacusis subjects using the previous literature.

2. Material and methods

2.1. Study design and population

Subjects participants were from an ongoing longitudinal study of age-related hearing loss (Dubno *et al.*, 2013), initiated in 1987 and conducted at the Medical University of South Carolina (MUSC). Measurements of auditory function, medical and noise exposure history, and DNA information were collected from all subjects. Subjects returned annually to receive an audiogram and speech recognition measures and to update medical history. The paired compositional dataset used in this paper was obtained by applying a supervised machine learning algorithm to audiograms from each ear for all subjects, as previously described (Dubno *et al.*, 2013). This dataset consists of 532 cases (subjects) with a total of 8 variables indicating the probabilities that a subject belongs to each of its four presbyacusis subtypes (Older-Normal, Metabolic, Sensory, and Metabolic+Sensory) in each ear. Detailed description of the dataset and calculation of these subtype probabilities can be found in Dubno *et al.* (2013).

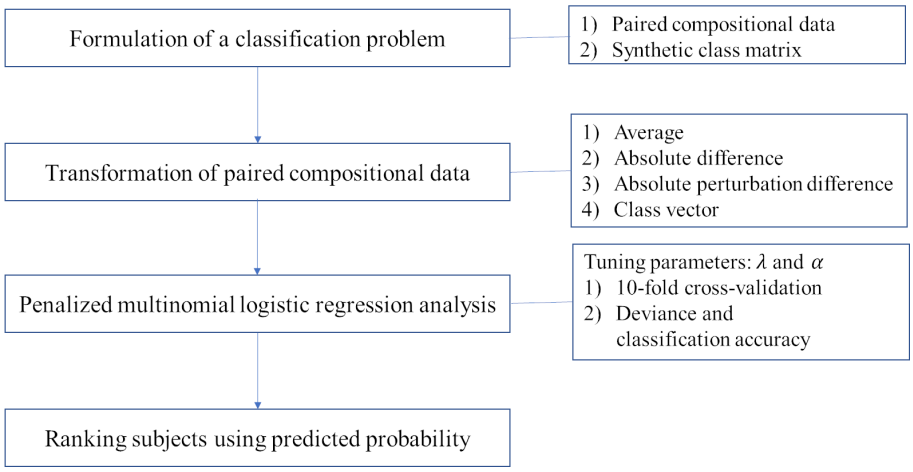

Figure 1: *Summary of the proposed workflow for ranking subjects based on paired compositional data.*

2.2. Notations

The paired compositional data with D components (subtypes) and n samples is expressed as a matrix of size $n \times 2D$, where $\mathbf{X} = \{(x_{ird}); i = 1, \ldots, n, r = 1, 2, d = 1, \ldots, D\}$, where x_{ird} represents the proportion of d^{th} component in the r^{th} ear in the pair (left or right) of i^{th} sample. Each ear corresponds to single compositional data; therefore, we have the constraint that $\sum_{d=1}^{D} x_{ird} = 1$ for $r = 1, 2$. Next, to formulate the ranking analysis as a supervised learning problem, we define the "synthetic class vector" $\mathbf{y_i} = (y_{i1}, y_{i2})$ by assigning each ear to the subtype with maximal probability (i.e., $y_{ir} = \arg\max_d x_{ird}$ for $r = 1, 2$). Thus, the paired compositional data $\mathbf{X}$ and the class matrix $\mathbf{y}$ can be written as

$$\mathbf{X} = \begin{pmatrix} x_{111} & \ldots & x_{11D} & x_{121} & \ldots & x_{12D} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n11} & \ldots & x_{n1D} & x_{n21} & \ldots & x_{n2D} \end{pmatrix}_{n \times 2D} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix}_{n \times 2}.$$

Figure 1 shows the overall workflow of the proposed supervised learning approach to rank subjects.

2.3. Transformation of paired compositional data

We generate three types of variables using paired compositional data in order to keep the rationale of 'phenotype purity' into account and facilitate interpretation. Here, the phenotype of a subject is defined "purer" if 1) the probability for one presbyacusis subtype is higher than those for other subtypes and 2) these probabilities are more comparable between two ears. First, we calculate the averaged probabilities for each component across two ears:

$$a_{id} = \frac{x_{i1d} + x_{i2d}}{2}, \quad i = 1, \ldots, n, \; d = 1, \ldots, D. \tag{2.1}$$

Second, we calculate the absolute difference in probabilities for each component between two ears:

$$b_{id} = |x_{i1d} - x_{i2d}|, \quad i = 1, \ldots, n, \; d = 1, \ldots, D. \tag{2.2}$$

Here, a_{id} and b_{id} measure relationships between two ears, given the component of compositional data. In addition to a_{id} and b_{id}, we also generate one more variable to measure informativeness within each

compositional data. For this purpose, we first apply the perturbation operator (Van den Boogaart and Tolosana-Delgado, 2013) to the compositional data from two ears:

$$f_{id} = \frac{x_{i1d}x_{i2d}}{\sum_{d'=1}^{D} x_{i1d'}x_{i2d'}}, \quad i = 1,\ldots,n,\ d = 1,\ldots,D. \tag{2.3}$$

Then, we calculate the absolute difference between the two largest values of the perturbation probabilities f_{id}:

$$h_i = \left|f_{i(D)} - f_{i(D-1)}\right|, \quad i = 1,\ldots,n,\ d = 1,\ldots,D, \tag{2.4}$$

where $f_{i(d)}$ means the d^{th} value in the ordered f_{id}, $d = 1,\ldots,D$. Note that each $(f_{i1},\ldots,f_{iD})$ is still located on a simplex because we employed the perturbation operator. Therefore, h_i allows us to measure how close i^{th} subject is to its closest vertex compared to its second closest vertex on the simplex. In this sense, h_i measures the degree of concentration in composition.

2.4. Synthetic class variable

For the modelling purpose, we also reformulate the synthetic class variable ($\mathbf{y}$). Note that subjects can be decomposed into two groups, including subjects that have the same subtype in both ears (pure) and those with different subtypes between two ears (impure). Based on this rationale, we generated a new $(D+1)$-level class variable z_i such that $z_i = d$ if $y_{i1} = y_{i2} = d$ for $d = 1,\ldots,D$ and $z_i = (D+1)$ if $y_{i1} \neq y_{i2}$. Thus, the transformed paired compositional data $\mathbf{W}$ and the modified class variable $\mathbf{z}$ can be written as follows.

$$\mathbf{W} = \begin{pmatrix} a_{11} & \ldots & a_{1D} & b_{11} & \ldots & b_{1D} & h_1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \ldots & a_{nD} & b_{n1} & \ldots & b_{nD} & h_n \end{pmatrix}_{n\times(2D+1)} \quad \text{and} \quad \mathbf{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}_{n\times 1},$$

where $z_i \in \{1,\ldots,(D+1)\}$, $i = 1,\ldots,n$.

It is of interest to consider a multivariate approach instead of transforming $\mathbf{y}$ to $\mathbf{z}$ because $\mathbf{y}$ is multivariate by nature. However, the transformation of $\mathbf{y}$ to $\mathbf{z}$ is still preferred due to the following two reasons. First, here our main goal is to identify subjects with purer phenotypes and utilize them for the EDP design. However, standard multivariate approaches are not tuned for this task and remain suboptimal for our problem. In contrast, this transformation is specifically designed with this goal in mind and is more efficient in distinguishing subjects with purer phenotypes. Second, we note that the two columns of $\mathbf{y}$ (y_{i1} and y_{i2}) are categorical variables. Although $\mathbf{y}$ itself is multivariate by nature, the choice of potential multivariate analysis approaches is rather limited because $\mathbf{y}$ is not continuous.

Instead of considering all the mismatches as one class ($z_i = (D+1)$ if $y_{i1} \neq y_{i2}$), we can also consider each of the mismatching pairs as a separate class. However, it is not of our main interest to distinguish different types of mismatches because our main goal is to identify subjects with purer phenotypes so that we can utilize them for the EDP design. Given this, considering different forms of mismatches as different classes can complicate interpretation. In addition, we also found that considering different forms of mismatches as different classes results in many classes with extremely small sample sizes. For example, there are only 5 and 2 subjects with the mismatches of (Metabolic, Normal) and (Met+Sen, Normal), respectively. Many classes with extremely small sample sizes can result in model fitting instability and generate other issues related to the imbalanced class problem. This can potentially be addressed by aggregating some of these classes; however, we again need to introduce another *ad hoc* procedure for this purpose, which might make interpretation harder. Based on this rationale, we decided to use the proposed approach.

2.5. Penalized multinomial logistic regression

While the data transformation procedures described in Sections 2.3 and 2.4 provide useful representations of the original paired compositional data and their class assignments, correlations among these transformed variables ($\mathbf{W}$) can still exist and interpretation of the results might not be straightforward. To address these issues, we use the penalized multinomial logistic regression model with the Elastic Net penalty function (Zou and Hastie, 2005).

When the number of classes is larger than two ($(D+1) > 2$), the traditional multi-logit model is defined as follows.

$$\log \frac{\Pr(Z = d \mid \mathbf{w})}{\Pr(Z = D+1 \mid \mathbf{w})} = \beta_{0d} + \mathbf{w}^T \boldsymbol{\beta}_d, \quad d = 1, \ldots, D, \tag{2.5}$$

where $z = (D+1)$ is set to the reference class. Alternatively, we can directly model the probability of each category (Zhu and Hastie, 2004) and in this case, the probability model is defined as

$$\Pr(Z = d \mid \mathbf{w}) = \frac{\exp\left(\beta_{0d} + \mathbf{w}^T \boldsymbol{\beta}_d\right)}{\sum_{d'=1}^{D+1} \exp\left(\beta_{0d'} + \mathbf{w}^T \boldsymbol{\beta}_{d'}\right)}, \quad d = 1, \ldots, D. \tag{2.6}$$

Based on this model, Friedman et al. modelled the penalized multinomial logistic regression using the regularized maximum multinomial likelihood (Friedman *et al.*, 2010), which we utilize here to fit the model. Let $z_{id} = I(z_i = d)$, $d = 1, \ldots, (D+1)$. Using the penalized log-likelihood, we can estimate β_{0d} and $\boldsymbol{\beta_d}$ by maximizing

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{d=1}^{D+1} z_{id} \log \Pr(Z_i = d \mid \mathbf{w}_i) - \lambda \sum_{d=1}^{D+1} P_\alpha(\boldsymbol{\beta}_d), \tag{2.7}$$

where $P_\alpha(\boldsymbol{\beta}_d)$ is a penalty term and λ and α are tuning parameters.

To handle correlations among the transformed variables ($\mathbf{W}$) and promote interpretation of coefficient estimates, we utilize the Elastic Net function for the penalty term $P_\alpha(\boldsymbol{\beta}_d)$. By combining the Lasso penalty (Tibshirani, 1996) with the Ridge penalty (Hoerl and Kennard, 1970), the Elastic Net penalty (Zou and Hastie, 2005) promotes sparsity in coefficient estimates and addresses correlation among predictors simultaneously. The Elastic Net function is defined as

$$P_\alpha(\boldsymbol{\beta}_d) = (1-\alpha)\frac{1}{2} \left\|\boldsymbol{\beta}_d\right\|_2^2 + \alpha \left\|\boldsymbol{\beta}_d\right\|_1, \tag{2.8}$$

where α is the tuning parameter that controls the balance between Lasso (L_1-norm) with Ridge (L_2-norm). Combining all together, the Elastic Net penalized negative log-likelihood becomes

$$\begin{aligned} l(\beta_{0d}, \boldsymbol{\beta}_d) = &- \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{d=1}^{D+1} z_{id} \left(\beta_{0d} + \mathbf{w}_i^T \boldsymbol{\beta}_d\right) - \log\left(\sum_{d=1}^{D+1} \exp\left(\beta_{0d} + \mathbf{w}_i^T \boldsymbol{\beta}_d\right)\right)\right\}\right] \\ &+ \lambda \sum_{d=1}^{D+1} \left((1-\alpha)\frac{1}{2} \left\|\boldsymbol{\beta}_d\right\|_2^2 + \alpha \left\|\boldsymbol{\beta}_d\right\|_1 \right). \end{aligned} \tag{2.9}$$

2.6. Parameter tuning

We use the 10-fold cross-validation to simultaneously select optimal values of the tuning parameters λ and α using accuracy and deviance as criteria. First, for each specific α, we identify the optimal λ value

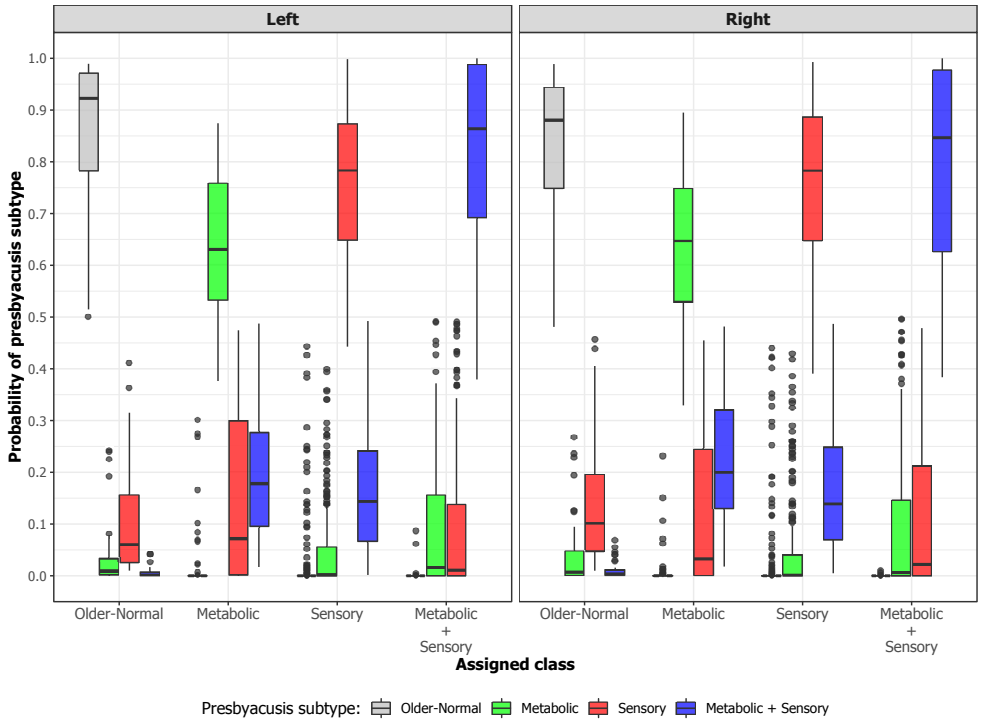

Figure 2: *Boxplots of probabilities for the four presbyacusis subtypes (color) for each assigned class (column), separated by ear (panel).*

corresponding to the minimum deviance and computed classification accuracy with this pair of λ and α. Then, we choose optimal λ and α with the minimum deviance corresponding to the maximum classification accuracy. We utilize the R package 'glmnet' (https://cran.r-project.org/web/packages/glmnet/) for parameter tuning and model fitting (Friedman *et al.*, 2010).

2.7. Subject ranking

After fitting the penalized multinomial logistic regression model that includes transformed paired compositional data and the modified synthetic class variable, we compute predicted probabilities for each subtype as $\Pr(z_i = d|\mathbf{w}_i)$, $d = 1, \ldots, D$. To evaluate phenotype purity, we then rank each subject based on predicted probabilities within each subtype.

3. Results

3.1. Data description

As described in Section 2.4, we defined the synthetic classes (y_{ir}) by assigning each ear to the subtype with maximal probability to reformulate the ranking problem as a supervised learning problem. Based on this definition, 379 subjects out of 532 have the same subtype in both ears (Older-Normal: 43, Metabolic: 32, Sensory: 173, Metabolic+Sensory: 131) while the remaining 153 subjects have different subtypes for their two ears. Therefore, 379 subjects belong to the "pure phenotype" classes

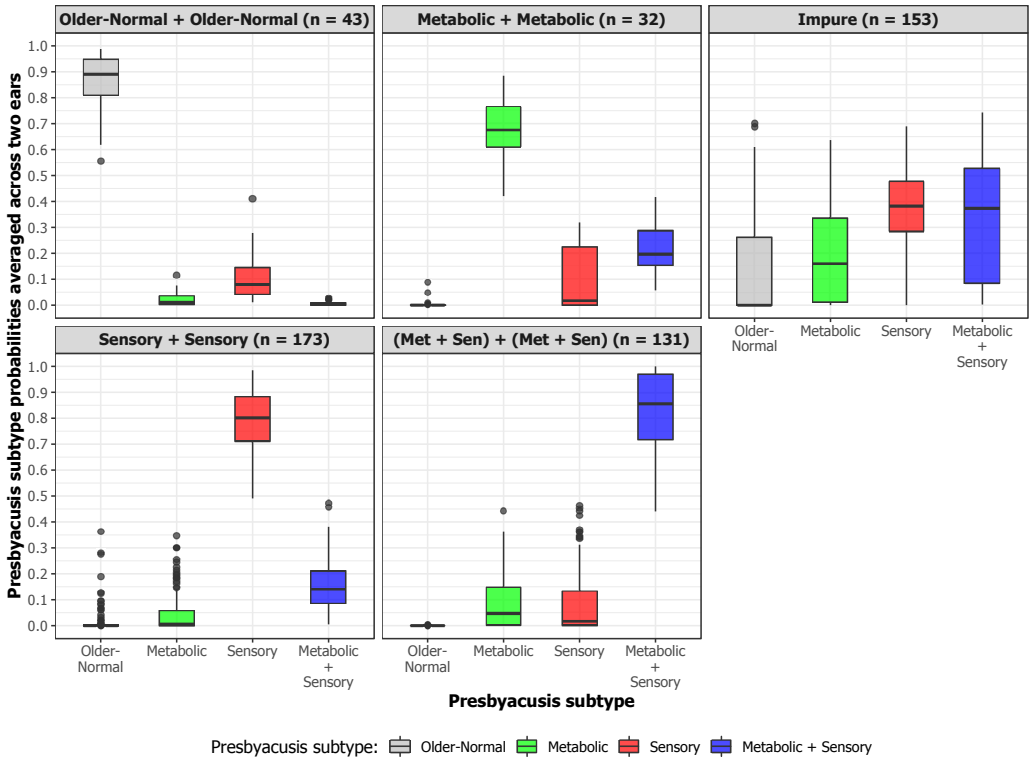

Figure 3: *Boxplots of probabilities averaged across two ears for all four presbyacusis subtypes (color), for four pure phenotype classes (cases in which the assigned classes were the same for the right and left ears, including Older-Normal, Metabolic, Sensory, and Metabolic+Sensory) and one impure phenotype class (cases in which assigned classes for the two ears were different) (panel).*

while the remaining 153 subjects belong to the "impure phenotype" class. Based on Figure 2 and Table A.1 in the Supplementary Materials, we notice that subtype probabilities corresponding to assigned classes in each ear are significantly higher than those for the other subtypes in each class. However, the difference in probabilities between Metabolic and other subtypes is smaller for the Metabolic class than other classes.

3.2. Exploratory data analysis

Figure 3 and Table A.2 in the Supplementary Materials show the averaged probabilities of each of four presbyacusis subtypes across two ears (a_{id}; Equation (2.1)) for four "pure phenotype" classes and one "impure phenotype" class (as described in Section 2.1). When subtypes are pure, the averaged probabilities for the class that subjects are assigned to are close to one, whereas the probabilities for other subtypes are close to zero. Note that the averaged probabilities for Metabolic+Sensory is higher when subjects were assigned to the Metabolic or Sensory classes for both ears, which might indicate that boundaries between Metabolic+Sensory and Metabolic subtypes as well as between Metabolic+Sensory and Sensory subtypes are less clear. When subtypes are impure, the averaged probabilities are similar between subtypes although these values are somewhat higher for Sensory

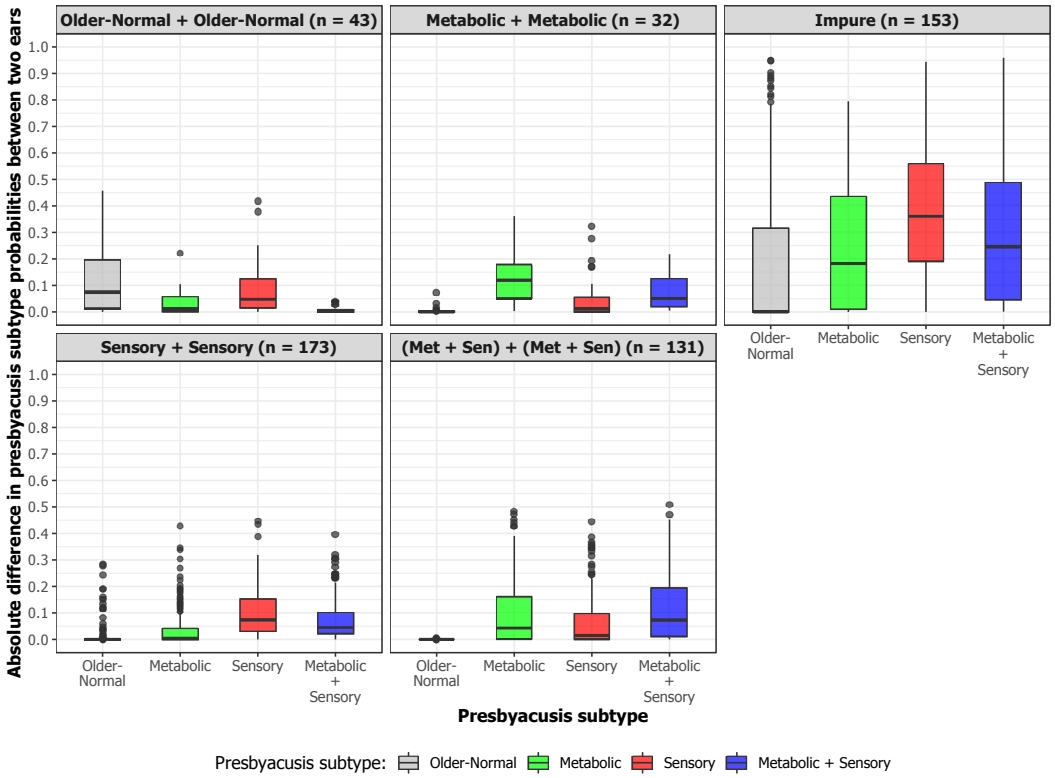

Figure 4: *Boxplots of absolute differences in presbyacusis subtype probabilities between two ears (color), for four pure phenotype classes (cases in which the assigned classes were the same for the right and left ears, including Older-Normal, Metabolic, Sensory, and Metabolic+Sensory) and one impure phenotype class (cases in which assigned classes for the two ears were different) (panel).*

and Metabolic+Sensory subtypes. Figure A.1 and Table A.3 in the Supplementary Materials further decomposed this impure phenotype class into all possible combinations of classes assigned to each of two ears. We observe that the averaged probabilities for the two classes that subjects are assigned to are closer to 0.5 when subtypes are impure.

Figure 4 and Table A.4 in the Supplementary Materials shows absolute differences in probabilities between two ears for each of the four presbyacusis subtypes (b_{id}; Equation (2.2)) for four pure phenotype classes and one impure phenotype class. We observe that these variables have values close to zero when subjects have pure phenotype classes. When subjects have the impure phenotype class, absolute difference values are overall higher than the pure phenotype classes. Figure A.2 and Table A.5 in the Supplementary Materials further decomposed this impure phenotype class into all possible combinations of classes assigned to each of two ears. When subjects have different classes for two ears, these variables have much higher values for many combinations and sometimes these values are even close to one.

Figure 5 and Table A.6 in the Supplementary Materials show perturbation probabilities for each of the four presbyacusis subtypes (f_{id}; Equation (2.3)), along with absolute differences between the two largest perturbation probabilities (h_i; Equation (2.4)), for four pure phenotype classes and one impure

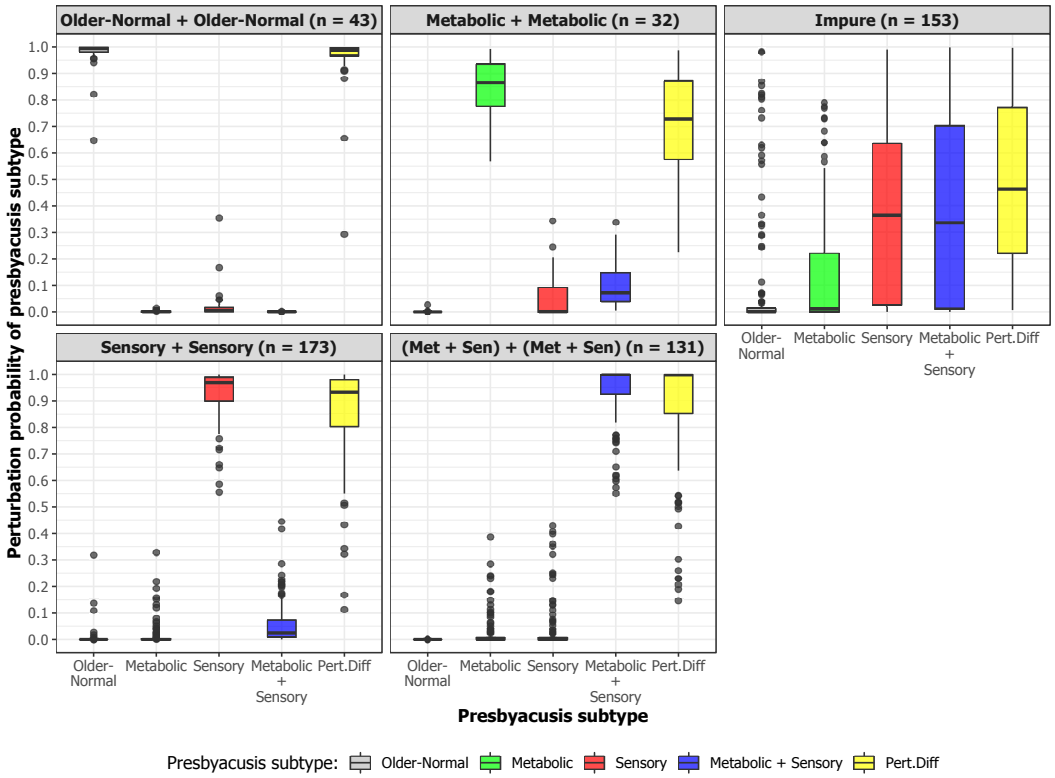

Figure 5: *Boxplots of perturbation probabilities for the four presbyacusis subtypes (the first four columns) and the absolute differences between the two largest perturbation probabilities ("Pert. Diff"; the last column) for four pure phenotype classes (cases in which the assigned classes were the same for the right and left ears, including Older-Normal, Metabolic, Sensory, and Metabolic+Sensory) and one impure phenotype class (cases in which assigned classes for the two ears were different) (panel).*

phenotype class. As in the case of averaged probabilities, when subtype is pure, the perturbation probability for the class that subjects are assigned to is close to one while those for other subtypes are close to zero. Similarly, when subtypes are impure, the perturbation probabilities for Sensory and Metabolic+Sensory subtypes are higher compared to Older-Normal and Metabolic subtypes. This is similar to what is observed for average probabilities, but differences are magnified here. The absolute differences between the two largest perturbation probabilities (h_i; Equation (2.4)) show interesting patterns and these values for the pure phenotype classes are significantly higher than those for the impure phenotype class.

In summary, data transformation results indicate that the three types of variables generated (a_{id}, b_{id}, h_i; Equations (2.1), (2.2), (2.4)) can provide information that can potentially be useful to predict the purity of phenotypes. Moreover, these three types of variables provide complementary information. The averaged probability (a_{id}) contains information about the degree of purity while the absolute difference in probabilities between two ears (b_{id}) indicates consistency in purity between two ears. Finally, the absolute differences between the two largest perturbation probabilities (h_i) represents the degree of concentration in composition.

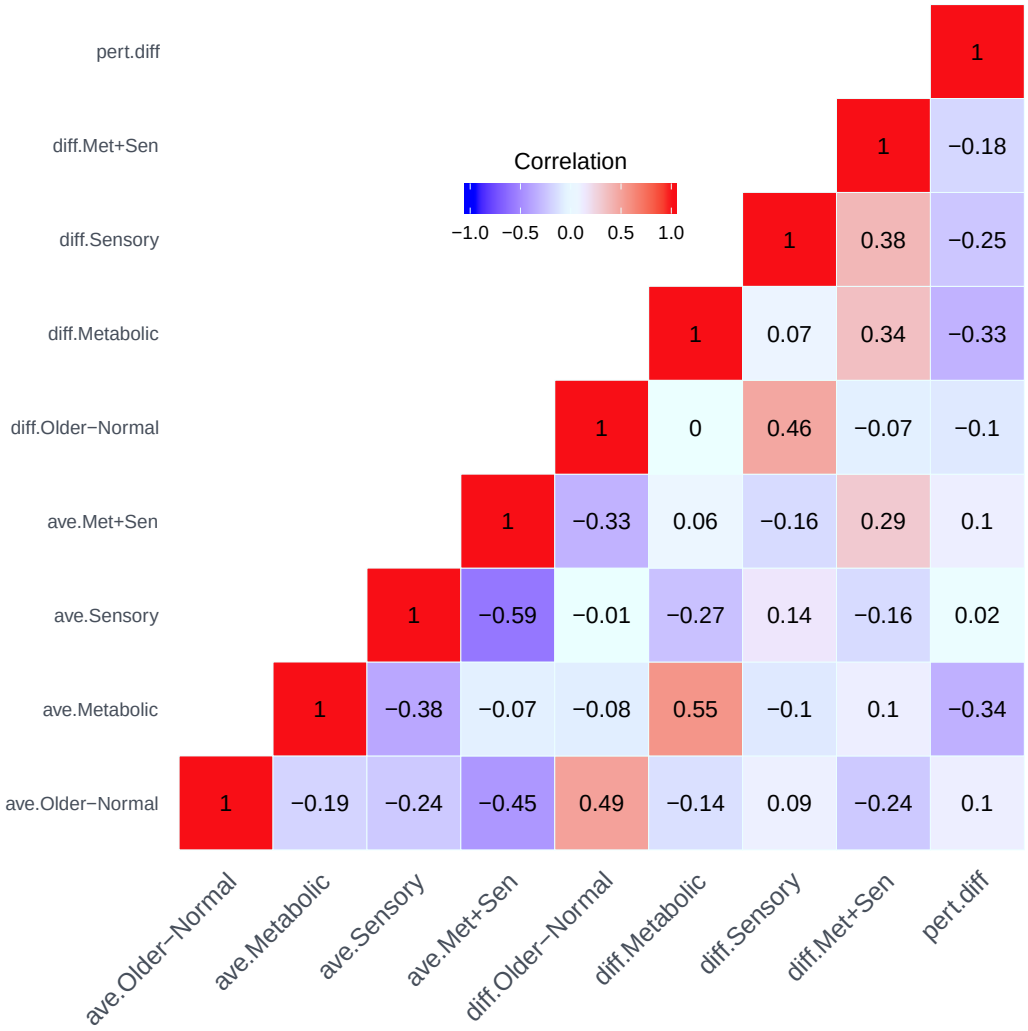

Figure 6: *Heatmap of correlation coefficients between three types of generated variables: averages (ave.*), absolute differences (diff.*), and absolute perturbation difference (pert.diff).*

3.3. Transformation of paired compositional data

Based on the exploratory analysis results in Section 3.2, we transformed the original 8 variables of paired compositional data (probabilities for 4 presbyacusis subtypes in each ear; $\mathbf{X}$) into 9 variables, including 4 subtype probabilities averaged between two ears (a_{id}, $d = 1,\ldots,4$; Equation (2.1)), 4 absolute differences in subtype probabilities between two ears (b_{id}, $d = 1,\ldots,4$; Equation (2.2)), and one absolute difference between the two largest values in the perturbation probabilities (h_i; Equation (2.4)). These 9 variables were used as predictors ($\mathbf{W}$). We call these three types of variables as averages, absolute differences, and absolute perturbation differences. In addition, we defined the response variable ($\mathbf{z}$) as categorical with 5 levels, including 4 classes with pure subtypes (Older-

Table 1: Coefficient estimates of the multinomial logistic regression model with the Elastic Net penalty

	Older-Normal	Metabolic	Sensory	Met+Sen	Impure
Intercept	−0.589	−1.768	−0.420	0.769	−2.009
ave.Older-Normal	8.285	0.000	−3.726	−2.330	−0.000
ave.Metabolic	−2.090	12.012	−4.656	−3.960	−0.646
ave.Sensory	−1.027	−0.913	9.456	−4.284	−0.000
ave.Meta+Sen	−1.429	−1.210	−3.931	9.010	−0.000
diff.Older-Normal	−3.122	0.000	−1.308	−1.320	−9.907
diff.Metabolic	0.000	−1.926	−1.058	−1.130	−8.944
diff.Sensory	−2.931	1.417	−9.506	0.000	11.388
diff.Met+Sen	−1.404	−0.981	0.000	−6.728	13.886
pert.diff	1.344	−0.179	4.608	2.244	−9.342

Ave.*, diff.*, and pert.diff mean the average, absolute difference, and absolute perturbation difference variables, respectively.

Normal, Metabolic, Sensory, and Metabolic+Sensory) and one class with impure subtype.

We begin by exploring the correlation among the 9 predictor variables ($\mathbf{W}$) that we generated (Figure 6). First, we observe that there are some positive correlations between average and absolute difference variables of the same subtype, especially for Metabolic (0.55) and Older-Normal (0.49) subtypes. Second, there are some negative correlations among average variables, e.g., between Sensory and Metabolic+Sensory subtypes (−0.59) and between Older-Normal and Metabolic+Sensory subtypes (−0.45). Third, there are some positive correlations among absolute difference variables, e.g., between Older-Normal and Sensory subtypes (0.46). Finally, the absolute perturbation difference is not strongly correlated with any other predictors.

We further checked the variance inflation factors (VIFs) for the transformed variables to investigate the multicollinearity between them. VIF values were obtained as follows. Ave.Older-normal = 6,967,878; Ave.Metabolic = 33,530,720; Ave.Sensory = 111,326,100; Ave.Met+Sen = 76,892,700; Diff.Older-normal = 1.71; Diff.Metabolic = 8.20; Diff.Sensory = 5.07; Diff.Met+Sen = 3.33; Pert.diff = 3.58. Essentially, high VIF values were observed for average variables while low VIF values were observed for the absolute difference and the absolute perturbation difference variables. Overall, this is consistent with what we observed in Figure 6.

3.4. Penalized multinomial logistic regression analysis

The final values of λ and α were determined as 0.0007 and 0.275, respectively (Figures A.3 in the Supplementary Materials). Hence, the estimated regression coefficients were only weakly shrunken compared to the original regression coefficients. Instead, correlations among predictors were weighted more heavily compared to the sparseness, which is reasonable given the strong within-dependency among components within the composition.

Table 1 shows regression coefficient estimates for the five classes we considered ($\mathbf{z}$). For each of the four pure phenotype classes, we observe that only the average variable corresponding to the class has positive coefficient estimate, while the average variables corresponding to other classes have negative coefficient estimates. This coincides with what is expected for compositional data because of its sum-to-one constraint ($\sum_{d=1}^{D} x_{ird} = 1$ for $r = 1, 2$). Most absolute difference variables have negative coefficient estimates and again the absolute difference variable corresponding to each class has the largest absolute coefficient estimate. Finally, the absolute perturbation difference variables have positive coefficient estimates for all pure phenotype classes except Metabolic. The results indicate the usefulness of the variables generated in our approach. Most of average variables are not selected in the case of impure phenotype class; however, all the absolute difference variables and the absolute

perturbation difference variable are selected. This suggests that averages and absolute differences are main drivers for the prediction of pure phenotype classes (Older-Normal, Metabolic, Sensory, and Metabolic+Sensory) while absolute differences mainly characterize the impure phenotype class. This is expected because the impure phenotype class is defined by presbyacusis subtype inconsistencies between ears, which are reflected more in the absolute difference variables. In contrast, the pure phenotype classes require high subtype probabilities in both ears, which are better reflected in the average variables.

To further explore sparseness and shrinkage of the proposed approach, we investigated two extreme cases of the Elastic Net penalty, which are the Lasso penalty ($\alpha = 1$) and Ridge penalty ($\alpha = 0$). The coefficient estimates of the Ridge model (Table A.7 in the Supplementary Materials) are similar to those of the Elastic Net model because our Elastic Net model is only weakly shrunken. However, compared to the Elastic Net model, the Ridge model put more weight on the average variables. In the case of the Lasso model (Table A.8 in the Supplementary Materials), the coefficient estimates are very sparse and only the average variable corresponding to each class is selected for all pure phenotype classes. In addition, the absolute difference variables corresponding to each subtype are also selected for Sensory and Metabolic+Sensory subtypes. Again, in the case of impure phenotype class, none of the average variables are selected, while all the absolute difference variables are included along with the absolute perturbation difference variable. These results confirm the findings from the Elastic Net model.

3.5. Subject ranking

We next evaluated patterns of predicted probability used to rank subjects, with respect to average, absolute difference and absolute perturbation difference variables (Figures A.4 in the Supplementary Materials). The results indicate that the predicted probabilities are positively associated with averages and absolute perturbation differences and negatively associated with absolute differences. Finally, we investigated our ranking results using the expected demographics based on previous literature (Dubno *et al.*, 2013). It has been reported that: 1) Older-Normal subtypes are more likely to be younger, female, and with a negative noise exposure history; 2) Metabolic subtypes are more likely to be older, female, and with a negative noise exposure history; 3) Sensory subtypes are more likely to be younger, male, with a positive noise exposure history; and 4) Metabolic+Sensory subtypes are more likely to be older, male, with a positive noise exposure history. Our ranking approach predicts phenotype purity for each presbyacusis subtype; therefore, it is more likely for the top-ranking subjects to have the demographics expected for each presbyacusis subtype along with demographics that are also more distinct among different presbyacusis subtypes. In contrast, for the lower ranking subjects, such separation among presbyacusis subtypes is expected to be more diluted. Based on this rationale, we partitioned the subjects for each pure phenotype class into the top-ranked 50% and the bottom-ranked 50% subjects using our ranking approach and compared distributions of demographics between the two groups.

Table 2 and Figures A.5–A.7 in the Supplementary Materials indicate that we observe the expected demographics for the top-ranked 50% of subjects (age, sex, and noise exposure history). First, Metabolic (mean = 75.69 years) and Metabolic+Sensory (mean = 74.98 years) subtypes are on average older than Older-Normal (mean = 67.12 years) and Sensory (mean = 70.57 years) subtypes. Second, the percentage of females is lower in Sensory (36.05%) and Metabolic+Sensory (51.52%) subtypes, compared to Older-Normal (86.36%) and Metabolic (68.75%) subtypes. Third, the percentage of subjects with a positive noise exposure history is higher in Sensory (60.47%) and Metabolic+Sensory (48.48%) subtypes, compared to Older-Normal (33.33%) and Metabolic

Table 2: Demographics (age, sex, and noise exposure history) of the top-ranked 50% and the bottom-ranked 50% subjects identified using the proposed statistical approach

Pure Phenotype Class	Age, Mean (SD)			Female, %			Positive Noise Exposure, %		
	All	Top 50%	Bottom 50%	All	Top 50%	Bottom 50%	All	Top 50%	Bottom 50%
Older-Normal (n = 43)	66.60 (6.82)	67.12 (6.97)	66.05 (6.80)	83.72	86.36	80.95	30.23	27.27	33.33
Metabolic (n = 32)	74.54 (6.91)	75.69 (7.42)	73.39 (6.38)	65.63	68.75	62.50	43.75	43.75	43.75
Sensory (n = 173)	71.50 (6.98)	70.57 (6.71)	72.42 (7.16)	43.93	36.05	51.72	53.49	60.47	46.51
Metabolic+Sensory (n = 131)	74.94 (7.39)	74.98 (7.48)	74.89 (7.36)	46.56	51.52	41.54	51.54	48.48	54.69
Overall (n = 379)	72.39 (7.54)			51.19			50.40		

(43.75%) subtypes. In contrast, differences among presbyacusis subtypes are more diluted for the bottom-ranked 50% of subjects as expected. For the bottom-ranked 50% subjects, standard deviations of means of age decrease from 4.01 to 3.89 compared to the top-ranked 50% subjects. Standard deviations of female percentages decrease from 21.72 to 16.85; in addition, standard deviations of positive noise exposure history percentages also decrease from 13.75 to 8.82. These results are consistent with what have been reported in the literature and further confirm the validity of our ranking approach.

4. Conclusion

In this paper, motivated by the problem to identify and rank subjects according to subtypes of age-related hearing loss, we proposed a new supervised learning approach to rank subjects using the penalized multinomial logistic regression model integrated with the transformation of compositional data. We formulated the problem as a supervised learning problem and utilized the penalized multinomial logistic regression model with the Elastic Net penalty, following transformation of the paired compositional data into three types of variables, including averaged probabilities of each component across two sides of the pair, absolute difference in component probabilities between two sides of the pair, and absolute difference between the two largest values in the perturbation probabilities. Its application to age-related hearing loss data indicates that the proposed supervised learning approach is effective for ranking subjects based on paired compositional data for age-related hearing loss subtyping. Furthermore, the results generated using our approach nicely coincide with biological knowledge and previous literature.

The proposed approach is promising for the subject ranking problem discussed in this manuscript; however, the literature for the analysis of paired compositional data is still limited to make more general recommendations at this point. There exists an important need to develop more statistical approaches to handle paired compositional data in various contexts. In general, there are multiple issues that need to be considered when analysis approaches are developed for the paired compositional data. First, if the original space is considered, it is important to take into account the 'sum-to-one' constraint carefully (such as visualizing data using ternary diagrams) that depicts observations on a simplex (Van den Boogaart and Tolosana-Delgado, 2013). Second, for the same reason, various statistical assumptions need to be carefully checked and addressed. While the popular ilr, clr, and alr transformations are known to help address these issues, assumption violations are still often observed such as the issue of heteroscedasticity (Maier, 2014). Finally, as we illustrated in our approach for the subject ranking problem, we need to consider additional and relevant issues, such as correspondence of each element

between the pair and relationships between them, when we analyze paired compositional data. We plan to investigate paired compositional data for diverse problems and develop statistical approaches to handle them in a future study.

Competing interests

The authors declare no conflict of interest.

Acknowledgement

This work was supported by the National Institute on Deafness and Other Communication Disorders [grant number P50-DC000422]; the National Institute of General Medical Sciences [grant number R01-GM122078]; the National Cancer Institute [grant number R21-CA209848]; the National Institute on Drug Abuse [grant number U01-DA045300]; and the National Institute of Arthritis and Musculoskeletal and Skin Diseases [grant number P30-AR072582]; and the South Carolina Clinical and Translational Research Institute with an academic home at the Medical University of South Carolina [National Center for Advancing Translational Sciences grant number UL1-TR001450].

References

Aitchison J (1982). The statistical analysis of compositional data, *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**, 139–160.

Barnett IJ, Lee S, and Lin X (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies, *Genetic Epidemiology*, **37**, 142–151.

Dubno JR, Eckert MA, Lee FS, Matthews LJ, and Schmiedt RA (2013). Classifying human audiometric phenotypes of age-related hearing loss from animal models, *Journal of the Association for Research in Otolaryngology*, **14**, 687–701.

Friedman J, Hastie T, and Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, **33**, 1–22.

Hoerl AE and Kennard RW (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

Huang BE and Lin D (2007). Efficient association mapping of quantitative trait loci with selective genotyping, *American Journal of Human Genetics*, **80**, 567–576.

Li D, Lewinger JP, Gauderman WJ, Murcray CE, and Conti D (2011). Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies, *Genetic Epidemiology*, **35**, 790–799.

Lin FR, Niparko JK, and Ferrucci L (2011). Hearing loss prevalence in the United States, *Archives of Internal Medicine*, **171**, 1851–1852.

Maier MJ (2014). DirichletReg: Dirichlet regression for compositional data in R, Research Report Series / Department of Statistics and Mathematics, 125, WU Vienna University of Economics and Business, Vienna.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

Vaden KI, Matthews LJ, Eckert MA, and Dubno JR (2017). Longitudinal changes in audiometric phenotypes of age-related hearing loss, *Journal of the Association for Research in Otolaryngology*, **18**, 371–385.

Van den Boogaart KG and Tolosana-Delgado R (2013). *Analyzing Compositional Data with R*,

Springer, Heidelberg.

Zhang G, Nebert DW, Chakraborty R, and Jin L (2006). Statistical power of association using the extreme discordant phenotype design, *Pharmacogenetics and Genomics*, **16**, 401–413.

Zhu J and Hastie T (2004). Classification of gene microarrays by penalized logistic regression, *Biostatistics*, **5**, 427–443.

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.

Received October 19, 2019; Revised December 11, 2019; Accepted December 20, 2019