JASA Express Letters

APRIL 01 2026

Objective comparison of audiometric profile frameworks across large-scale datasets

Chen Xu

JASA Express Lett. 6, 044402 (2026)
https://doi.org/10.1121/10.0043212

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Objective comparison of audiometric profile frameworks across large-scale datasets

Chen Xu[1,2,a)]
[1]Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, 26111 Oldenburg, Germany
[2]Institute of Sound and Vibration Research, Faculty of Engineering and Physical Sciences, University of Southampton, Southampton, United Kingdom

Abstract: Audiometric profiles classify individuals according to patterns of hearing loss derived from the audiogram. Although several audiogram-based profiling frameworks have been proposed, the influence of dataset characteristics on their structural performance has not been systematically examined. This study compared six established audiometric profiling frameworks across five large-scale datasets from the United States and Germany using the Davies-Bouldin score and principal component analysis. Clustering performance based on the Davies-Bouldin score was largely comparable across datasets, although profile-specific differences were observed. These findings inform the robustness and generalizability of audiogram-based classification frameworks across large-scale samples.

[Editor: Qian-Jie Fu] https://doi.org/10.1121/10.0043212

Received: 29 January 2026 **Accepted:** 13 March 2026 **Published Online:** 1 April 2026

1. Introduction

Audiometric profiles refer to classification frameworks that group individuals according to patterns of hearing loss derived from the audiogram. Such frameworks provide a structured way to summarize heterogeneous audiometric configurations, reduce dimensionality, and facilitate comparison across studies and populations. Multiple audiogram-based profiling frameworks have been proposed (e.g., Bisgaard *et al.*, 2010; Cruickshanks *et al.*, 2020; Dubno *et al.*, 2013; Humes, 2019, 2021; Parthasarathy *et al.*, 2020). Understanding the robustness of audiometric profiling across datasets is important for assessing their generalizability and interpretability in large-scale research contexts. The present study evaluates whether these classification frameworks produce clearly distinguishable groups across multiple epidemiological cohorts.

Despite the widespread use of the audiometric profiles, the factors influencing their generation are not yet fully understood. Our previous studies (Xu *et al.*, 2025; Xu *et al.*, 2026) investigated two such factors—the number of profiles and the clustering approach—and found that both significantly affected audiometric profile generation. However, other factors, such as the choice of dataset for bench-marking, may also have a substantial impact. In Xu *et al.* (2026), we used a clinical dataset [the extended Oldenburg Hearing Health Record (OHHR)] (Jafri *et al.*, 2025) that included the results of both routine audiological tests and numerous other tests. Since these additional tests are typically not used in standard clinical practice, it remains unclear whether our previous findings generalize to large-scale datasets containing only routine clinical measures. Furthermore, the sample in our earlier dataset was highly targeted, consisting predominantly of older adults with hearing impairment, thus representing a relatively narrow demographic. This raises the question of how well these findings generalize beyond the analyzed datasets. To address these limitations and uncertainties, the present study aims to compare different audiometric profiling approaches using large-scale datasets.

The National Health and Nutrition Examination Survey (NHANES) is a U.S. national database designed to assess the health and nutrition of adults and children (Humes, 2023a). It includes participants across all age groups and with diverse types of hearing loss, making it a valuable large-scale epidemiological dataset for structural analyses. Audiometric assessments in NHANES consist of four components: (1) pre-exam audiometric questions, (2) otoscopy, (3) middle-ear testing, and (4) pure-tone air-conduction audiometry. For the air-conduction audiogram, thresholds for both ears were measured at seven frequencies (0.5–8 kHz) using a modified Hughson–Westlake procedure with an audiometer. In the present study, the three most recent NHANES databases [2011–2012 ($N = 3818$), 2015–2016 ($N = 4263$), and 2017–2020 ($N = 3837$)] were analyzed to derive audiometric profiles. The year labels (e.g., 2011–2012) indicate the periods of data collection.

A German large-scale dataset was also included (von Gablenz *et al.*, 2020). This dataset comprises air-conduction audiograms collected in Emden and Oldenburg (Northern Germany, 2010–2012) and in Aalen (Southern Germany, 2008–2009), with a total of $N = 3105$ participants. Hearing thresholds were assessed at eight frequencies (0.25–8 kHz) using

a)Corresponding author: chen.xu@uni-oldenburg.de

manual audiometry. Compared with the smaller OHHR dataset ($N = 1127$ participants), these large-scale datasets provide a broader age range and greater diversity of audiograms.

In our previous study, eight auditory profile frameworks were evaluated on the OHHR database with respect to cluster quality; that is, the extent to which participants within the same cluster were similar and those in different clusters were distinct. The baseline audiometric profiles classified participants into two groups based on the pure-tone average across four frequencies (PTA4; mean threshold at 0.5, 1, 2, and 4 kHz): normal hearing (PTA4 < 20 dB HL) and hearing impaired (PTA4 ≥ 20 dB HL). The Bisgaard profiles (Bisgaard *et al.*, 2010) provided a more fine-grained classification into ten groups, comprising seven N-type audiograms (flat or moderately sloping) and three S-type audiograms (steeply sloping). The World Health Organization (WHO) hearing impairment (HI) grades categorized participants into six levels—normal, mild, moderate, moderately severe, severe, and profound—based on the audiogram (Humes, 2019). Similarly, the Wisconsin Age-Related Hearing Impairment Classification Scale (WARHICS) divided hearing-impaired participants into eight categories (Humes, 2019, 2021). With regard to audiometric phenotypes, Dubno *et al.* (2013) proposed five types—older-normal, pre-metabolic, metabolic, sensory, and combined metabolic and sensory. Building on this, Parthasarathy *et al.* (2020) employed a data-driven Gaussian mixture model to derive four profiles, later referred to as general phenotypes: normal audiogram, flat sloping hearing loss, high-frequency hearing loss, and mixed sensorineural hearing loss. A detailed comparison of these frameworks in terms of cluster quality was reported in Xu *et al.* (2025) and Xu *et al.* (2026). However, it remains unclear how these audiometric profile frameworks perform in terms of clustering quality when applied to large-scale and more diverse datasets.

Taken together, this study addresses the following research questions:

1. RQ1: How do different audiometric profile frameworks perform in terms of clustering quality for the large-scale datasets?
2. RQ2: To what extent does the choice of dataset influence the generation of audiometric profiles, and are the results consistent across the large-scale datasets?
3. RQ3: Are the findings from the large-scale datasets consistent with those obtained from the clinical OHHR dataset?

2. Methods

2.1 The large-scale datasets

In this study, we focus on the single test applied across the large-scale datasets, namely, air-conduction pure-tone audiometry. The supplementary material summarizes the methodological differences in audiogram assessments between the German large-scale dataset and the NHANES dataset. Notably, the procedures are largely comparable due to adherence to the standardized guidelines of the National Center for Health Statistics. The supplementary material details the test equipment, software, and test personnel. All devices were calibrated, and measurements were conducted in soundproof booths compliant with ISO standards. Furthermore, the measurement procedures for all datasets followed ISO 8253-1. Thus, although certain components differed (e.g., choice of audiometer), the overall quality of audiogram measurements across datasets can be regarded as well-controlled and standardized. For the NHANES datasets, raw (unweighted) audiometric records were analyzed. Sampling weights were not applied, as the primary aim was to evaluate patterns of audiometric profiles rather than estimate population prevalence.

The NHANES 2017–2020 dataset was divided into two age-defined subsets—Older ($\geq$70 years) and Youth (6–19 years)—and all analyses were performed separately within each subgroup. Table 1 presents demographic characteristics for the five datasets. Most datasets included more than 2800 participants, with the exception of the NHANES 2017–2020 Older cohort, which comprised approximately 1000 adults. The NHANES 2017–2020 Older cohort included the oldest participants, whereas the NHANES 2017–2020 Youth database included the youngest. The mean ages of the NHANES 2011–2012 and NHANES 2015–2016 cohorts were approximately 44 years. All five datasets were sex-balanced.

2.2 Audiometric profile frameworks

Participants from all five datasets were categorized according to six audiometric profiling frameworks: the baseline audiometric profile, the Bisgaard profile, the WHO HI grades, the WARHICS levels, the audiometric phenotype, and the general phenotype, based on the respective work of Bisgaard *et al.* (2010), Humes (2019), Cruickshanks *et al.* (2020), Dubno *et al.* (2013), and

Table 1. Demographic information for the five datasets. N = number of participants.

	German	NHANES 2011–2012	NHANES 2015–2016	NHANES 2017–2020 Older	NHANES 2017–2020 Youth
N	3105	3818	4263	946	2891
Age (years)	54.6 (±17.5)	43.6 (±14.4)	44.3 (±14.2)	75.4 (±3.7)	12.8 (±3.8)
Sex, *n*	Male:1445; Female: 1660	Male: 1944; Female: 1874	Male: 2036; Female: 2227	Male: 435; Female: 511	Male: 1439; Female: 1452

Parthasarathy *et al.* (2020). Comprehensive auditory profiles (e.g., Saak *et al.*, 2022) could not be generated, as the datasets lacked supra-threshold measures. The supplementary material shows the number of participants assigned to each audiometric profile for the six audiometric profiling frameworks and five datasets. Some participants were not assigned to some audiometric profiles for the audiometric phenotype and general phenotype frameworks.

2.3 Data analysis

First, the Davies–Bouldin (DB) score was applied to assess the cluster quality of different audiometric profiling frameworks. The DB score can be calculated as follows (Davies and Bouldin, 1979):

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} R_{ij}, \tag{1}$$

where R_{ij} denotes the similarity measure between the i_{th} group and the j_{th} group, with j corresponding to the group most similar to group i, defined as follows:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}, \tag{2}$$

where s_i represents the average distance between each point for group i and the centroid of that group (i.e., cluster diameter) while d_{ij} denotes the distance between group centroids i and j. A lower DB score indicates a better classification into clusters. The DB score has a lower bound of zero. To estimate the distribution of the DB scores, bootstrapping was performed by drawing 1000 samples with replacement from a given dataset, each containing $N = 1000$ data points. The DB score was calculated using the Python "scikit-learn" package. To ensure a fair comparison across audiometric profiling frameworks, we followed Xu *et al.* (2026) and normalized the DB scores with respect to the number of profiles by dividing the original score by $\log_2(n)$, where n denotes the number of profiles. Following the approach described in Humes (2023a), cumulative distribution functions were estimated using percentile values (5th–95th percentiles) of the normalized DB score, and Cohen's h was calculated across percentiles and averaged to provide an overall measure of practical significance. |h| values of 0.20, 0.50, and 0.80 correspond to small, medium, and large effect sizes, respectively.

Principal component analysis (PCA) was applied to the five datasets to enable a visual comparison of audiometric profiles (e.g., Bisgaard profiles and audiometric phenotypes). Feature contributions to the first and second principal components (PC1 and PC2) were examined following the approach described by Lê *et al.* (2008), allowing interpretation of the underlying dimensions captured by these components. PCA visualizations were generated using the "factoextra" package (Kassambara and Mundt, 2017), with PC1 and PC2 plotted on the x and y axes, respectively. Prior to PCA, all features were standardized to have unit variance. The final PCA model retained five dimensions.

3. Results

Figure 1 shows normalized DB scores for the six audiometric profiling frameworks and five datasets. A lower normalized DB score reflects both high between-group variance (i.e., participants with different profiles are well separated) and low within-group variance (i.e., participants with the same profile are closely grouped). Across all datasets, the Bisgaard profiles yielded the lowest normalized DB scores, followed by WHO HI grades and WARHICS levels. These profiles showed significantly lower DB scores than the baseline audiometric profile ($p < 0.05$). In contrast, the audiometric phenotype exhibited significantly higher DB scores than the baseline.

To quantify practical significance, effect sizes were estimated using average Cohen's |h| derived from percentile-based cumulative distribution functions. Results are reported in Table 2. Across datasets, the Bisgaard profiles showed very small effect sizes relative to the baseline ($|h| = 0.03$–0.14), indicating that distributional differences in normalized DB scores were minimal despite statistical significance. Similarly, WHO HI grades and WARHICS levels were associated with trivial-to-small effect sizes ($|h| \leq 0.16$), suggesting limited practical differences in clustering performance. In contrast, the audiometric phenotype yielded consistently larger effect sizes ($|h| = 0.13$–0.34), reflecting small-to-moderate distributional deviations and supporting its poorer clustering performance.

Overall, the Bisgaard profiles showed the best statistical clustering performance across datasets, although differences were small. The audiometric phenotype exhibited consistently higher DB scores and larger effect sizes, reflecting poorer clustering performance.

Figure S1 in the supplementary material shows feature contributions of the first two principal components (PCs) for each dataset. For the German dataset [Figs. S1(A) and S1(B) in the supplementary material], PC1 was primarily determined by PTA4 and hearing thresholds at medium frequencies (e.g., 1 and 2 kHz), whereas PC2 was primarily determined by thresholds at the frequency extremes (e.g., 0.25 and 8 kHz). For the NHANES 2011–2012 and 2015–2016 datasets, the main contributors to PC1 were PTA4 and thresholds at 0.75 kHz, while PC2 was determined by thresholds at 0.25 and 0.5 kHz, reflecting the influence of low-frequency hearing. A similar pattern was observed for NHANES 2017–2020 Older and Youth, where medium frequencies contributed most to PC1, and extreme frequencies contributed most to PC2.

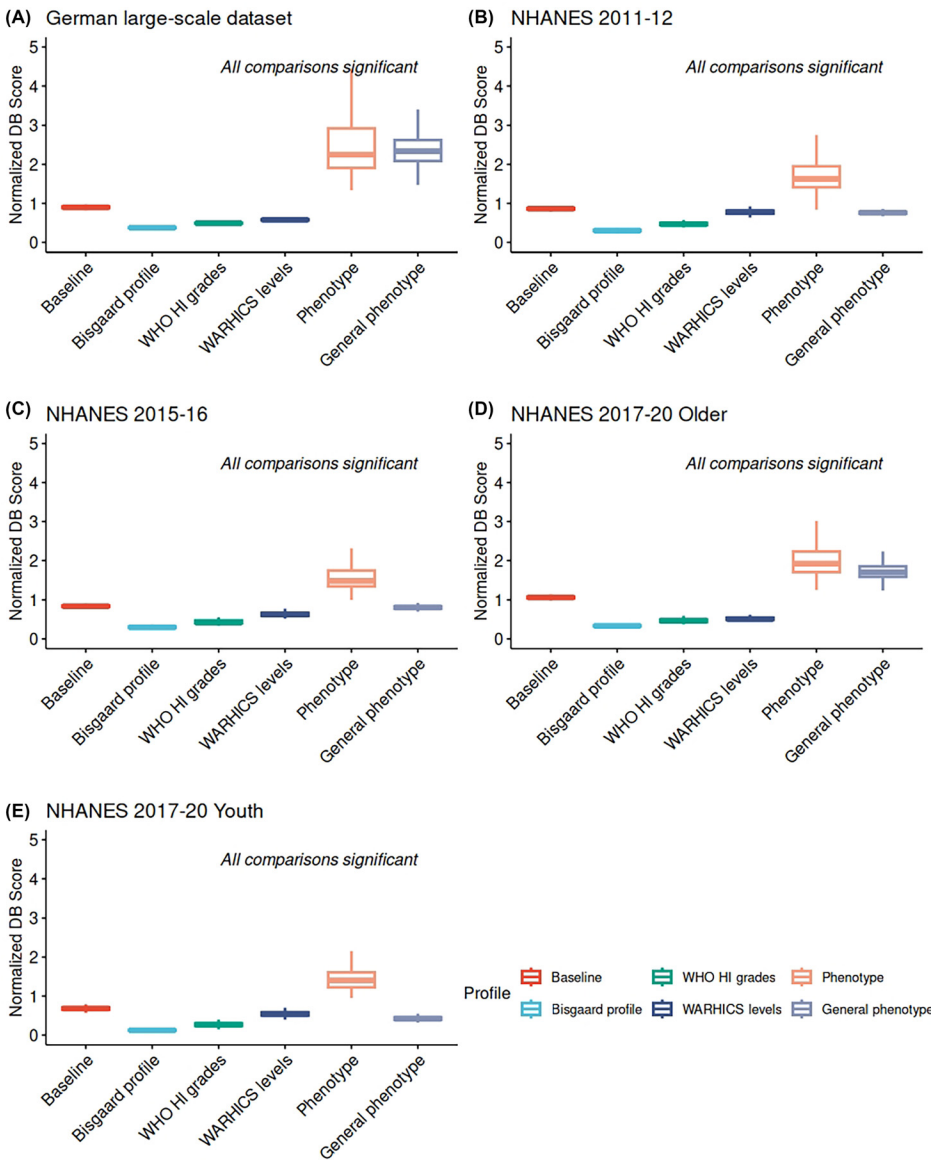

Fig. 1. Normalized DB scores for six audiometric profiling frameworks across five datasets: (A) German large-scale dataset, (B) NHANES 2011–2012, (C) NHANES 2015–2016, and (D) NHANES 2017–2020 Older, and (E) NHANES 2017–2020 Youth. A lower normalized DB score indicates better clustering performance. See the supplementary material for the definitions of the audiometric profiling frameworks. The boxplots show the median, 25th and 75th percentiles, and interquartile range. Whiskers extend to the most extreme values within 1.5 × interquartile range from the lower and upper quartiles. Statistical comparisons between audiometric profiling frameworks were conducted using pair-wise *t* tests with Bonferroni correction.

Table 2. Average Cohen's |h| values quantifying distributional differences in normalized DB scores relative to the baseline audiometric profile. Effect sizes were derived from percentile-fitted cumulative distribution functions (5th–95th percentiles). Conventional benchmarks (Cohen, 1988): small (0.2), medium (0.5), large (0.8).

Framework	German	NHANES 2011–2012	NHANES 2015–2016	NHANES 2017–2020 Older	NHANES 2017–2020 Youth
Bisgaard	0.03	0.05	0.07	0.14	0.05
WHO HI grades	0.06	0.04	0.13	0.04	0.14
WARHICS	0.04	0.07	0.10	0.12	0.16
Audiometric phenotype	0.27	0.13	0.34	0.16	0.21
General phenotype	0.14	0.01	0.03	0.07	0.19

Overall, PC1 primarily reflects hearing thresholds at medium frequencies, whereas PC2 captures variability at the low- and high-frequency ends.

Our results differ from those reported in previous studies. For example, Encina-Llamas *et al.* (2025) identified the PTA4 as the main contributor to PC1 and the audiogram slope as the primary contributor to PC2. In contrast, Sanchez-Lopez *et al.* (2020) and Wu *et al.* (2022) reported that PC1 was mainly driven by low-frequency hearing thresholds, whereas PC2 was dominated by high-frequency thresholds. These discrepancies are both attributable to differences in the datasets used and to variations in the included feature sets, sample characteristics, and measurement protocols. Specifically, some studies considered additional derived features, such as slope indices, while others restricted the analysis to audiometric thresholds; the analyzed samples also differed in terms of age distribution and hearing loss severity, which can shape the variance structure of the data; and differences in audiometric equipment and calibration across studies may have introduced systematic shifts. Together, these factors contribute to the lack of consistent interpretations of PC1 and PC2 across studies.

Figures 2(A)–2(E) illustrate PCA for each dataset with participants classified by Bisgaard profiles. For all five datasets, PC1 accounted for the majority of variance, reflecting the dominant role of overall hearing level in structuring the data. PC2 explained an additional 10.6%–19.1% of the variance, capturing a secondary but meaningful dimension related to audiogram configuration. Together, PC1 and PC2 explained about 80%–90% of the variability, indicating that most of the heterogeneity across Bisgaard profiles can be represented by two dimensions.

The separation observed in Fig. 2 further supports this interpretation. N-type audiograms (red) and S-type audiograms (blue) formed distinct clusters, suggesting that PC1 and PC2 together reliably differentiate profiles according to both degree of hearing loss and audiogram slope. Within the N-types, clear distinctions were visible for profiles N1–N4, while within the S-types, profiles S1–S3 were also well separated. Thus, PCA not only reduces the dimensionality of the data effectively but also preserves clinically meaningful differences between Bisgaard profiles.

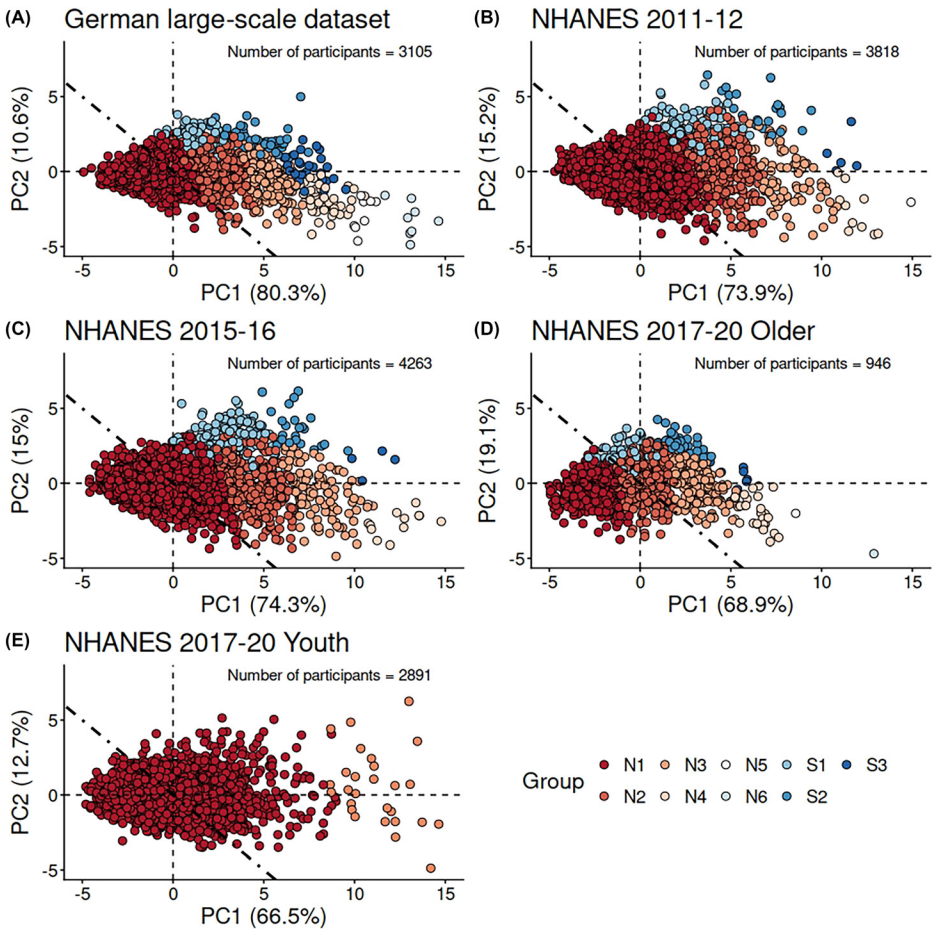

Fig. 2. PCA of the five datasets based on Bisgaard profiles. The *x* axis represents PC1, and the *y* axis represents PC2. Participants with different Bisgaard profiles are shown by different colors, with the number of participants indicated in the top-right corner. Class N7 is absent, as no participants with this profile occurred in the cohort.

Comparing Figs. 2(A)–2(E), the PCA results for the German large-scale dataset, NHANES 2011–2012, NHANES 2015–2016, and NHANES 2017–2020 Older were largely consistent, whereas the NHANES 2017–2020 Youth dataset showed notable differences. The NHANES 2017–2020 Youth cohort predominantly comprises younger participants (see Table 1). Consistent with this, audiometric profiles representing more severe degrees of hearing loss are largely absent in this cohort (see the supplementary material). This distributional difference likely contributes to the deviations relative to the other four datasets. Furthermore, some discrepancies were observed between the German dataset and the U.S.-based NHANES datasets, which is in line with findings of von Gablenz *et al.* (2020) who reported systematic differences in pure-tone thresholds between European and U.S. cohorts. Accordingly, variations in the PCA results across datasets are expected, given the heterogeneity in audiometric characteristics.

Figure S2 presents the PCA results based on audiometric phenotypes. A large proportion of participants could not be classified into any phenotype, a limitation noted previously (Saak *et al.*, 2025). Also, noticeable overlaps were observed between phenotypes, which may reflect shared underlying auditory or metabolic characteristics. For instance, participants with "Older-normal" and "Pre-metabolic" profiles frequently overlapped, suggesting a potential continuum rather than distinct categories. Comparing Figs. S2(A)–S2(E), the findings align with earlier observations: the PCA results for the NHANES 2017–2020 Youth differed markedly from those for the other four datasets, and the German large-scale dataset showed systematic differences compared to the U.S.-based NHANES datasets. Possible explanations for these discrepancies were discussed above.

4. Discussion

4.1 Consistency across large-scale datasets

Overall, the cluster quality results, as measured by normalized DB scores, were consistent across the five datasets. This suggests that the normalized DB score is a robust indicator for quantitatively comparing different audiometric profiling approaches, largely independent of factors, such as sample demographics or age distribution. Previous studies (e.g., Elkhouly *et al.*, 2021; Xu *et al.*, 2026) have also demonstrated the reliability of the DB score for evaluating audiometric profiles. Notably, the Bisgaard profiles consistently achieved good DB scores across all datasets. These profiles were originally derived from large-scale empirical data and subsequently refined and validated by experts. As a result, they are more constrained than the other audiometric profile frameworks and align with clinically meaningful categories of hearing loss. Because the prototypes have been "cleaned" and validated by experts, Bisgaard profiles exhibit less variability than frameworks based on raw audiograms without experts' refinements. Furthermore, their design emphasizes distinct categories (e.g., flat, sloping, steeply sloping), which enhances between-cluster separation—a key determinant of cluster quality. Bisgaard profiles can be regarded as prototype-based, clinically grounded, with expert-optimized prototype patterns aimed at practical applicability, which limit within-profile variability and enhance separability while retaining interpretability.

The PCAs were not consistent across the five datasets. One possible reason is the inherent statistical differences between the datasets (e.g., age, sex distribution, and hearing thresholds). In contrast to quantitative metrics, such as the DB score, PCA is affected by such dataset-specific characteristics.

4.2 Comparison between large-scale epidemiological and clinical datasets

Our previous studies (Xu *et al.*, 2025; Xu *et al.*, 2026) conducted similar comparisons using the OHHR. Comparing the findings of the current study with those of the previous ones, two points emerge.

First, the comparisons based on DB scores were consistent across both large-scale epidemiological and clinical datasets. Despite differences in sample demographics and variability in audiometric characteristics, both studies indicated that the Bisgaard profiles yielded good cluster quality, whereas the audiometric phenotypes performed poorly. The ranking of the audiometric profile frameworks was also stable: the highest cluster quality was observed for the Bisgaard profiles, followed by WHO HI grades, WARHICS levels, baseline profiles, general phenotypes, and finally audiometric phenotypes. These consistent results across datasets confirm that the DB score is a robust indicator, largely independent of study sample and test battery.

Second, substantial differences emerged when comparing the PCA results of the current study (Figs. 2 and S2) with those of the previous one [Fig. 2(B) for Bisgaard profiles and Fig. 2(F) for audiometric phenotypes]. These differences confirm that PCA is strongly affected by the inherent statistical properties of the dataset. In Xu *et al.* (2025), the test battery included not only audiograms but also supra-threshold measures, making the data more complex and increasing the difficulty of PCA computation. This complexity likely contributed to the poor visualization of group separations in the earlier study, where scores for participants with different Bisgaard profiles often overlapped, and participants within the same profile were not well clustered. By contrast, in the current study, which relied solely on audiogram data, group separation between audiometric profiles became more apparent, as reflected in the higher explained variance of the PCA.

4.3 Outlook and limitations

Both the previous study and the present one used the normalized DB score for quantifying cluster quality and applied PCA for visual comparisons. These tools have generally performed well across datasets with varying inherent statistics, and

they provide a useful basis for objective validation. Nevertheless, subjective evaluation of audiometric profile frameworks remains limited. Expert input could support the refinement of auditory profiling frameworks by ensuring that derived profiles are clinically interpretable and correspond to meaningful diagnostic and intervention-relevant distinctions. This need for incorporating expert perspectives has been emphasized in the literature (e.g., Dimitrov *et al.*, 2025).

Although auditory profiles can be derived efficiently with few parameters (Sanchez-Lopez *et al.*, 2020; Saak *et al.*, 2022) and further optimized in terms of the cluster quality (Xu *et al.*, 2025), only a few studies have addressed hearing-aid fitting strategies based on these optimized profiles (e.g., Sanchez-Lopez *et al.*, 2021; Cañete *et al.*, 2024). Developing novel fitting strategies tailored to refined and more specific auditory profiles therefore represents an important next step. For example, average measurement outcomes within each profile (e.g., audiograms, loudness functions, or speech recognition thresholds) could serve as inputs for fitting formulas.

Because NHANES sampling weights were not applied, demographic distributions in the analyzed sample may not fully reflect those of the U.S. population. Finally, due to the limitations of the available datasets, we were not able to compare more comprehensive auditory profiling approaches (e.g., the BEAR profiles, the Hearing4all profiles, as analyzed in Xu *et al.*, 2026), which incorporate both audiograms and supra-threshold measures. Future work should therefore focus on building datasets that include supra-threshold tests, potentially enabled by mobile-device–based assessments (see Xu *et al.*, 2024a, Xu *et al.*, 2024b).

Humes and Zapala (2024) introduced a three-digit triad audiogram-classification framework in which separate pure-tone averages for low-, mid-, and high-frequency regions are used to characterize hearing loss configuration. This approach provides a more frequency-specific representation than a single PTA4 measure and may better capture configuration-related variability. Although the present study did not evaluate this scheme, future work could incorporate the triad framework within the proposed evaluation framework to examine its clustering performance relative to existing systems.

5. Conclusion

Six audiometric profile frameworks were compared across five large-scale datasets using the normalized DB score to assess cluster quality and PCA for visual comparison. The normalized DB scores were consistent across datasets, with the Bisgaard profile achieving the highest score and the audiometric phenotype the lowest. PCA, applied to the Bisgaard and audiometric phenotype frameworks, revealed notable differences between datasets, indicating its high sensitivity to dataset-specific characteristics. The normalized DB score provides a robust and objective measure of cluster quality, whereas PCA enables visualization of dataset-specific differences in the multivariate profile structure. Together, these methods enable an objective comparison of audiometric profile frameworks, which may support the optimization of audiometric profiles in terms of the cluster quality and, in the future, provide a theoretical foundation for profile-based hearing device fitting.

Supplementary Material

In the supplementary material, Table S1 compares audiogram assessment methods between the German large-scale and NHANES datasets, and Table S2 summarizes participant distributions across the six audiometric profiling frameworks in the five datasets. In addition, Fig. S1 presents the feature contributions to the first (PC1) and second (PC2) principal components across the five datasets. Figure S2 shows the PCA of audiometric phenotypes across the five datasets.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy (EXC 2177/1, Project ID 390895286).

Author Declarations

Conflict of Interest

No potential conflict of interest was reported by the author.

Data Availability

The data that support the findings of this study are derived from publicly available sources. The NHANES audiometric datasets (2011–2012, 2015–2016, and 2017–2020) are available from the U.S. Centers for Disease Control and Prevention. The German large-scale audiometric data analyzed in this study have been published previously and are available as described in the cited references. Processed data and analysis scripts are available from the corresponding author upon reasonable request.

References

Bisgaard, N., Vlaming, M. S., and Dahlquist, M. (**2010**). "Standard audiograms for the IEC 60118-15 measurement procedure," Trends Amplif. **14**(2), 113–120.

Cañete, O. M., Loquet, G., Sánchez-López, R., Hougaard, D. D., Schnack-Petersen, R., Gaihede, M., and Neher, T. (**2024**). "Auditory profile-based hearing aid fitting: Self-reported benefit for first-time hearing aid users," Audiol. Res. **14**(1), 183–195.

Cohen, J. (**1988**). *Statistical Power Analysis for Behavioral Sciences*, 2nd ed. (Erlbaum, Mahwah, NJ).

Cruickshanks, K. J., Nondahl, D. M., Fischer, M. E., Schubert, C. R., and Tweed, T. S. (**2020**). "A novel method for classifying hearing impairment in epidemiological studies of aging: The Wisconsin age-related hearing impairment classification scale," Am. J. Audiol. **29**(1), 59–67.
Davies, D. L., and Bouldin, D. W. (**1979**). "A cluster separation measure," IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-1**(2), 224–227.
Dimitrov, L., Barrett, L., Chaudhry, A., Muzaffar, J., Lilaonitkul, W., and Mehta, N. (**2025**). "Uncovering phenotypes in sensorineural hearing loss: A systematic review of unsupervised machine learning approaches," Ear Hear. **46**, 1401–1411.
Dubno, J. R., Eckert, M. A., Lee, F. S., Matthews, L. J., and Schmiedt, R. A. (**2013**). "Classifying human audiometric phenotypes of age-related hearing loss from animal models," J. Assoc. Res. Otolaryngol. **14**, 687–701.
Elkhouly, A., Andrew, A. M., Rahim, H. A., Abdulaziz, N., Abdulmalek, M., Mohd Yasin, M. N., and Siddique, S. (**2021**). "A novel unsupervised spectral clustering for pure-tone audiograms towards hearing aid filter bank design and initial configurations," Appl. Sci. **12**(1), 298.
Encina-Llamas, G., Kjaerbùl, E., and Kressner, A. (**2025**). "Towards a hearing loss map: Phenotyping complex audiological data beyond audiometry," in *International Symposium on Auditory and Audiological Research*, Nyborg, Denmark.
Humes, L. (**2023a**). "Development and application of a reference-interval approach to tympanometric norms using U.S. population data for ages 6-80+ years," Am. J. Audiol. **32**(4), 908–929.
Humes, L. E. (**2019**). "The World Health Organization's hearing-impairment grading system: An evaluation for unaided communication in age-related hearing loss," Int. J. Audiol. **58**(1), 12–20.
Humes, L. E. (**2021**). "Further evaluation and application of the Wisconsin age-related hearing impairment classification system," Am. J. Audiol. **30**(2), 359–375.
Humes, L. E. (**2023b**). "Hearing thresholds for unscreened US adults: Data from the National Health and Nutrition Examination Survey, 2011–2012, 2015–2016, and 2017–2020," Trends Hear. **27**, 23312165231162727.
Humes, L. E., and Zapala, D. A. (**2024**). "Easy as 1-2-3: Development and evaluation of a simple yet valid audiogram-classification system," Trends Hear. **28**, 23312165241260041.
Jafri, S., Berg, D., Buhl, M., Vormann, M., Saak, S., Wagener, K. C., and Kollmeier, B. (**2025**). "The Oldenburg Hearing Health Record (OHHR)," Sci. Data **12**(1), 1546.
Kassambara, A., and Mundt, F. (**2017**). "factoextra: Extract and visualize the results of multivariate data analyses (R package version 1.0.7)," available at https://CRAN.R-project.org/package=factoextra.
Lê, S., Josse, J., and Husson, F. (**2008**). "FactoMineR: An R package for multivariate analysis," J. Stat. Softw. **25**, 1–18.
Parthasarathy, A., Romero Pinto, S., Lewis, R. M., Goedicke, W., and Polley, D. B. (**2020**). "Data-driven segmentation of audiometric phenotypes across a large clinical cohort," Sci. Rep. **10**(1), 6704.
Saak, S., Huelsmeier, D., Kollmeier, B., and Buhl, M. (**2022**). "A flexible data-driven audiological patient stratification method for deriving auditory profiles," Front. Neurol. **13**, 959582.
Saak, S., Oetting, D., Kollmeier, B., and Buhl, M. (**2025**). "Integrating audiological datasets via federated merging of auditory profiles," Trends Hear. **29**, 23312165251349617.
Sanchez-Lopez, R., Fereczkowski, M., Neher, T., Santurette, S., and Dau, T. (**2020**). "Robust data-driven auditory profiling towards precision audiology," Trends Hear. **24**, 2331216520973539.
Sanchez-Lopez, R., Fereczkowski, M., Santurette, S., Dau, T., and Neher, T. (**2021**). "Towards auditory profile-based hearing-aid fitting: Fitting rationale and pilot evaluation," Audiol. Res. **11**(1), 10–21.
von Gablenz, P., Hoffmann, E., and Holube, I. (**2020**). "Gender-specific hearing loss in German adults aged 18 to 84 years compared to US-American and current European studies," PLoS One **15**(4), e0231632.
Wu, M., Christiansen, S., Fereczkowski, M., and Neher, T. (**2022**). "Revisiting auditory profiling: Can cognitive factors improve the prediction of aided speech-in-noise outcome?," Trends Hear. **26**, 23312165221113889.
Xu, C., Hülsmeier, D., Buhl, M., and Kollmeier, B. (**2024a**). "How does inattention influence the robustness and efficiency of adaptive procedures in the context of psychoacoustic assessments via smartphone?," Trends Hear. **28**, 23312165241288051.
Xu, C., Kollmeier, B., and Schell-Majoor, L. (**2026**). Objective comparison of auditory profiles using manifold learning and intrinsic measures. arXiv:2601.03827.
Xu, C., Schell-Majoor, L., and Kollmeier, B. (**2024b**). "Development and verification of non-supervised smartphone-based methods for assessing pure-tone thresholds and loudness perception," Int. J. Audiol. (published online).
Xu, C., Schell-Majoor, L., and Kollmeier, B. (**2025**). "Optimizing auditory profiling for precision audiology: A comparative study of clustering frameworks," in *International Symposium on Auditory and Audiological Research*, Zenodo, Nyborg, Denmark.