PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

A clinical test battery for Better hEAring Rehabilitation (BEAR): Towards the prediction of individual auditory deficits and hearing-aid benefit

Raul H SANCHEZ-LOPEZ[1], Silje Grini NIELSEN[1], Oscar CAÑETE[1], Michal FERECZKOWSKI[1], Mengfan WU[2], Tobias NEHER[2], Torsten DAU[1] & Sébastien SANTURETTE[1,3].

[1] Hearing Systems Section, Dept. of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Institute of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark

[3] Oticon A/S, Smørum, Denmark

ABSTRACT

One aim of the Better hEAring Rehabilitation (BEAR) project is to define a new clinical profiling tool, a test battery, for individualized hearing loss characterization. Recently, Sanchez-Lopez et al. (2019) proposed a test battery that includes six types of measures: audibility, middle-ear analysis, speech perception, binaural-processing abilities, loudness perception, and spectro-temporal resolution. The results of 75 listeners were analyzed using a data-driven approach (Sanchez-Lopez et al., 2018), which provided evidence for the existence of two independent sources of distortion and four different auditory profiles. The classification of the listeners into auditory profiles allows the prediction of the performance of the listeners on different psychoacoustic tasks as well as their expected aided speech intelligibility. For clinical practice, a decision tree with a small set of highly predictive tests is desirable for an efficient classification of hearing-impaired individuals. The main aim of the present study was to investigate the optimal decision tree and to propose a clinically feasible test battery with a minimum number of tests for accurate listener classification. The clinical test battery will be used in a large-scale field study that will help implement a hearing-aid fitting protocol for better hearing rehabilitation.

Keywords: Audiology, data-driven, supra-threshold distortions

1 INTRODUCTION

The Better hEAring Rehabilitation (BEAR) project pursues the development and implementation of new methods for the diagnosis of hearing deficits as well as new hearing-aid compensation strategies to improve hearing rehabilitation. Since digital hearing aids were introduced to the market, hearing-aid users have reported increased benefit (1), probably because of the advanced signal processing techniques (or features) that are now commonly available, such as directionality, noise reduction and dynamic range compression. However, the hearing-aid fitting is still based on the audiogram only which provides the basis for frequency-dependent gain prescription. The other features are adjusted based on preferences and not according to the individual auditory deficits of the user. Furthermore, in hearing care clinics it is common to “fine-tune” some hearing-aid parameters during follow-up visits (2). If the initial fitting is near-optimal, the follow-up visit may focus on individualization of the fitting parameters according to the “life-style” of the patient. However, if the initial fitting is far from optimal, the audiologist needs to tailor-fit hearing-aid parameters to the hearing deficits of the listener by “trial-and-error”. The BEAR project attempts to improve this situation by identifying groups of listeners – or “auditory profiles” – with specific performance patterns on a range of threshold and supra-threshold tasks and by providing tailored solutions with proposed dedicated hearing-aid compensation strategies for each auditory profile.

In an attempt to identify the auditory profiles, Sanchez-Lopez et al. (3) hypothesized that the hearing deficits of a given listener can be described as the combination of two independent types of auditory distortions. The hypothesis was based on the idea that each type of distortion can cause both threshold and supra-threshold deficits and that these deficits are not necessarily independent. In Figure

1.a), the two types of distortions create a two-dimensional space where a given listener's location is determined by the degree of severity of these distortions. As a result, the listener can be identified as belonging to a certain auditory profile. As shown in Figure 1, normal-hearing listeners are located at the bottom left-hand corner, exhibiting no distortions. Profile A corresponds to a group with minor distortions and therefore good performance in general. Profile C exhibits a high degree of both types of distortions. Profile B exhibits a high degree of distortion type I. Profile D shows a high degree of distortion type II. Using a data-driven approach (3), four auditory profiles were identified by analyzing the data from two previous studies, providing evidence for the validity of this approach. However, the substantial differences in terms of listeners and tests applied in these two studies limited the overall conclusions that could be drawn from this work.

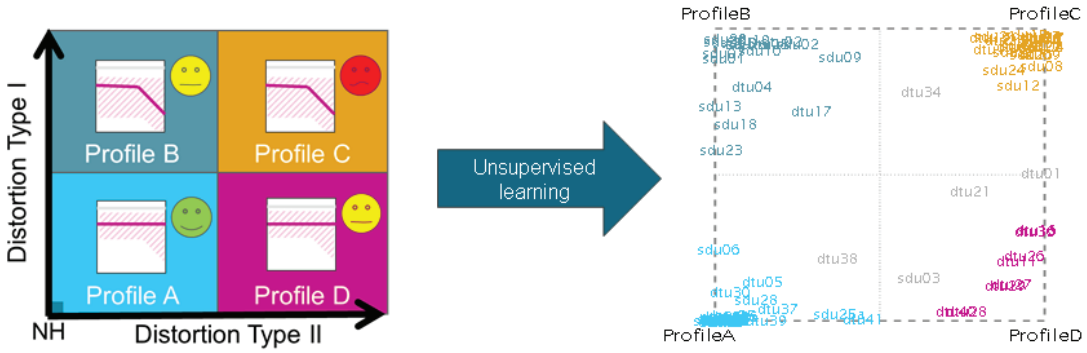

Figure 1 – Hypothesized auditory profiles together with the results of the data-driven profile identification. Left panel: The listeners are placed in a two-dimensional space along two dimensions of auditory distortion. Right panel: Using a data-driven analysis, listeners are placed in the two-dimensional space as a function of their probability of belonging to a specific profile.

In order to overcome the aforementioned limitations discussed in (3), a new test battery including a range of supra-threshold psychoacoustic tests was proposed and evaluated in 75 listeners with various types of audiometric configurations. Additionally, the test-retest reliability of the new test battery was investigated in a subset of 11 listeners (4). The dataset obtained in this manner will in the following be referred to as $BEAR_3$. For the classification of the 75 listeners, unsupervised learning techniques were used to carry out iterative auditory profiling based on the data-driven approach (5). After this analysis, 70 of the 75 listeners were reliably identified as belonging to one of the four auditory profiles A-D and the remaining five listeners (shown in grey in Figure 1) were left unclassified. However, this iterative unsupervised method requires the entire dataset to identify the four groups and is therefore not suitable for the classification of new listeners. Decision trees are a well-known simple classification tool that may prove useful for classifying unseen data, i.e. new listeners. The efficacy of decision trees can be explored by evaluating their classification performance (6).

When implementing a new protocol for diagnosing a specific disease in the clinic, it is crucial to evaluate its ability to correctly identify the patients who are affected by the health problem under consideration. In general, two types of errors can occur in this classification process: truly affected patients may be "missed" (false negatives) and healthy patients may be "misclassified" as being affected (false positives). Confusion matrices are typically used to quantify the test performance of a classifier. In addition to the classification performance, it is of interest to investigate the cost efficiency of a new clinical protocol (7) by estimating the cost of having false negatives or false positives as well as the benefit that the correct classification would provide.

The goal of the current study was to develop a decision tree for a large field study to be conducted as part of the BEAR project, where listeners will have to be classified into the four hypothesized auditory profiles. It is also of interest to identify an additional group of unclassified listeners (Uc) who do not seem to belong to any of the four primary profiles. The $BEAR_3$ dataset was used for

investigating the accuracy and efficiency of different decision trees. Using supervised learning techniques, different classification strategies were tested and evaluated in terms of both test performance and cost effectiveness.

2 Methods

Decision tree classifiers were trained for predicting the identified auditory profiles from (5) using supervised learning. The analysis of the cost efficiency was based on considerations made in the context of the aforementioned field study to be carried out in different hearing clinics. These considerations cover the recruitment of a random sample of 500 listeners where at least 60 listeners per profile are expected.

2.1 Classification methods

The classification of the listeners into the four auditory profiles was performed using supervised learning with tests that showed good to excellent reliability as input, and the labels of the four auditory profiles as well as the unclassified group (Uc) as output. The classification algorithm used here was a standard classification and regression tree (CART), which makes use of recursive binary partitions in order to fit the data to the best set of binary decisions or splits (8). Four classification schemes were considered:

- **DTA:** A simple classification based only on the audiogram.
- **DT10:** A multi-label "fitted" classification. Decision tree based on all reliable tests and 10 binary decisions.
- **DT7:** A multi-label "pruned" classification with seven binary decisions.
- **DT4:** A multi-label "pruned" classification with four binary decisions.

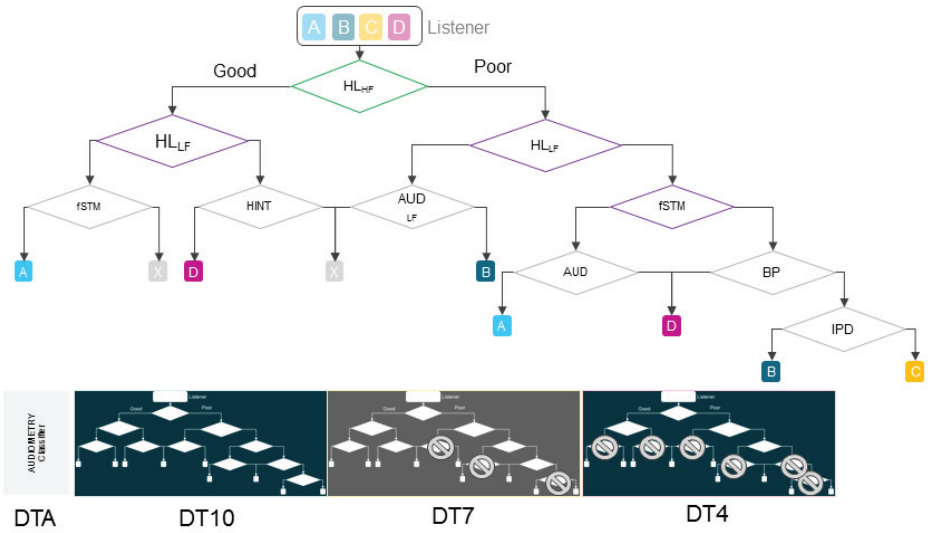

Figure 2 – Top panel: The complete decision tree (DT10). Bottom panel: The four decision trees considered in the present study. DTA: Audiometry-based classifier, DT10: Same as in top panel; DT7: Decision tree with seven binary decisions (DT10 with three pruned splits), and DT4: Decision tree with four binary decisions (DT10 with six pruned splits).

Figure 2 illustrates the complete classification tree (DT10). Each diamond (split) corresponds to a logic rule related to a given variable, for example HL_{HF} > 45 dB HL. The right branch corresponds to poorer outcome and the left branch to better outcome. The details of the logic rules are not shown here. The decision trees DT7 and DT4 are the result of pruning the decision tree DT10 by discarding some of the nodes, as illustrated in the figure.

2.2 Test performance and cost-efficiency evaluation

In order to evaluate both the classification performance and the cost efficiency, a dataset was created for bootstrapping. The original dataset was copied seven times which resulted in 525 observations. Next, the specific standard error of the measurements (SEM) of each test (4) was used for introducing some uncertainty (additive Gaussian noise) in the outcomes to simulate the data from the aforementioned field study. Confusion matrixes were then calculated for 100 iterations and the cost-efficiency was estimated.

The cost-efficiency was calculated based on (7) and adapted for the multi-label case. Consider a 2x2 matrix of costs **C.** Following the previous assumptions, C_{00} is the cost of a true negative, i.e. a participant to be excluded from the study or correctly "not-classified" as a given profile. The cost C_{00} would be equal to the session cost. C_{11} is the cost of a correct classification. C_{01} and C_{10} correspond to false positives and false negatives, respectively, which would introduce outliers in the final results. These would be equal to the cost of misclassification. Additionally, consider the matrix **P** with the probabilities of each of the previous cases, where P_{11} is the probability of correct classification, P_{00} that of correct rejection, and P_{01} and P_{10} those of the two types of misclassification. The expected cost is the Hadamard product of the **P** and **C** matrixes:

$$Expected\ Cost = \mathbf{P} \circ \mathbf{C} \tag{1}$$

This generic expression can then be simplified due to the fact that the probability of belonging to a given class $class_i$ not truly belonging to that group $P(class_i|class_j)$ is equal to $1 - P(class_j|class_j)$. The index i denotes the predicted class and the index j the actual class. Therefore, the expression can be simplified to:

$$Expected\ Cost = \sum_i \sum_j P(class_i|class_j) - \beta\, P(class_j|class_j)\,, \tag{2}$$

where β is defined as

$$\beta = \frac{C_{11} - C_{01}}{C_{00} - C_{10}} \frac{P(class_j)}{1 - P(class_j)}. \tag{3}$$

Given that the probabilities can be calculated in terms of the specificity and sensitivity, Equation 2 can be written as follows:

$$Expected\ Cost = \sum_j Specificity_j - \beta\, Sensitivity_j\,. \tag{4}$$

Equation 4 provides the extension of the expected cost as defined in (7) but for the multi-class case. It was used here to calculate the expected cost for each iteration to estimate the final expected cost per session.

2.2.1 Cost assumptions

The following assumptions were made:

- **Test cost:** Each additional test that is not part of current clinical practice incurs costs for the implementation, the training of the examiners and the documentation. Uus et al. (9) analyzed the costs of implementing a newborn screening program. The average set-up cost across 16 sites for implementing two tests was £665 for 1000 infants. Therefore, in the present study, a hypothetical total cost of $600 was considered for a field study that involves 500 listeners. The cost per new test per session would therefore be $1.2.
- **Session cost**: The duration of the session has a cost that involves the salary of the examiner and the use of the facilities. Taking the average of the costs suggested in (10–12) and assuming that one session lasts for one hour, this leads to a hypothetical cost of $60 per session or $1 per minute.
- **Correct classification**: The correct classification of a given listener increases the probability for the study to be successful. As suggested in (7), this should involve the long-term benefits, including the future reduction of follow-up visits in the clinics if the project is a success. In this case, we limited the expected benefit to the reduction of follow-up

visits. Tecca (13) recently studied the number of visits and the incidence of hearing-aid fitting-related complaints during the first six weeks of hearing aid use. It was shown that the first and second visit involved changes in the gain and advanced features in more than 70% of the cases. Since the new BEAR fitting rationale aims to provide a better first-fit solution, a cost of one follow-up visit ($60) was considered here.

- **Misclassification cost:** The cost of classifying the listener as belonging to a different auditory profile would correspond to the extra efforts used for this listener to obtain their optimal fitting, i.e. follow-up visits. Here, it is assumed that these listeners would be unsatisfied with their fittings because that corresponds to any other auditory profile and that this would lead to two extra follow-up visits for fine-tuning and verification ($120).

Table 1 shows the characteristics of each of the considered classifiers in terms of number of tests, the duration of the session and the total test cost and session cost.

Table 1- Description of the four classifiers in terms of number of tests, duration and costs. The number of tests includes the outcome measure HINT by default even in the case of DTA and DT4 where this test is not part of the decision trees.

Decision Trees	Description	Number of tests	Duration (min)	Test cost ($)	Session cost ($)
DTA	Audiometry classifier	1	27	1.2	28.2
DT10	Complete classifier	5	68	6	74
DT7	Pruned classifier I	4	41	4.8	45.8
DT4	Pruned classifier II	3	34	3.6	37.6

3 Results

3.1 Classification performance

The four classifiers were tested with a constructed data set based on the original data of the $BEAR_3$ data after applying bootstrapping. Figure 3 shows the confusion matrices of the four classifiers.

Figure 3 – Confusion matrices corresponding to one single iteration of classification for each of the four tested decision trees.

The audiometry-based classifier (DTA) was able to correctly predict 67% of the data with a low sensitivity in the predictions of profiles B and C and a low specificity in the case of profile D (19%). The classifiers DT10, DT7 and DT4 had an overall accuracy of 85%. However, they differed in terms of the specificity and sensitivity for each of the profiles. DT10 and DT7 were able to identify some Uc listeners correctly. In contrast, while DT4 lead to a higher specificity in the classification of profile D, it had the disadvantage that all the Uc listeners would be classified as any of the four profiles. The main difference between DT10 and DT7 was the sensitivity, especially for the Uc listeners. DT10 was more accurate for the Uc listeners, and it had also higher specificity for some of the other profiles. DT7 predicted less Uc listeners and misclassified more true B listeners, but had also higher sensitivity for profile C. Overall, the decision trees that contain binary decisions for identifying Uc listeners (DT10 and DT7) were both more accurate and more specific.

3.2 Expected cost

The expected cost was calculated according to Equation 4. Additionally, the total expected cost of the field study was estimated by the sum of the fixed costs, the planned sessions for 500 listeners recruited randomly, and the additional sessions needed for fulfilling the requirement of testing 60 listeners in each profile.

Figure 4 shows the expected costs. The left panel illustrates the differences among the four decision trees where DTA resulted in higher costs than the other three classifiers. DT10 provided higher cost-efficiency with $4 gained per listener, followed by DT7 and DT4 with $1.5 each per listener. The right panel of Figure 4 shows the total cost of the field study. The audiometry-based DTA classifier was the one with the lowest fixed and planned costs but the highest total cost. This is because of the risk of misclassification, which requires numerous additional listeners to get 60 subjects in profile D, with a total of 1992 listeners. DT10 and DT7 required a similar number of listeners (~675 listeners in total) but differed in the session cost, making DT7 a cheaper decision tree overall. The last decision tree DT4 provided the lowest total cost and required a lower number of additional measurements (564 listeners in total). However, the disadvantage is that DT4 cannot identify Uc listeners and would therefore classify them as belonging to one of the four auditory profiles. If the aim is to achieve a low number of misclassified listeners, this classifier would not be the optimal choice.

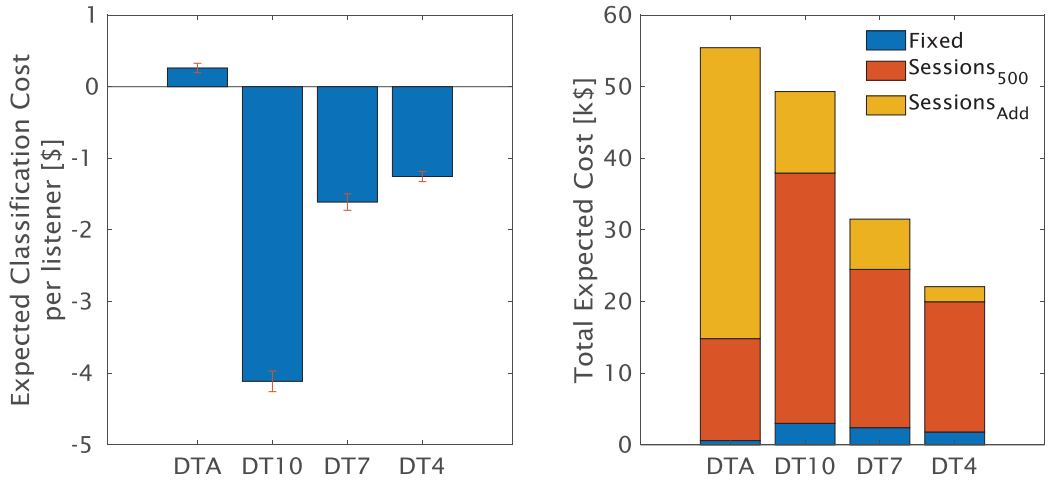

Figure 4 – Expected classification cost per session and total classification cost. The total classification cost involves the costs of the implementation of new tests (Fixed), the session costs for 500 listeners ($Sessions_{500}$) and the cost of additional sessions ($Sessions_{Add}$).

Overall, the results suggest that DT10 would be the best candidate for the considered field study, due to its higher sensitivity and specificity. Moreover, the clinical test battery that can help to better define the auditory profiles in a larger population by gathering information related to auditory spectro-temporal resolution, speech intelligibility, loudness perception and binaural processing abilities.

4 Discussion

The results of the present study speak in favor of a reduced test battery based on five tests included in DT10 for classifying listeners in clinical practice. These tests include adaptive categorical loudness scaling (ACALOS), the hearing in noise test (HINT), the binaural pitch (BP) test, the frequency threshold for identifying interaural phase differences (IPD), and a fast version of the spectro-temporal modulation sensitivity test (fSTM). As such, the proposed clinical version of the test battery covers four domains: loudness perception, speech-in-noise intelligibility, binaural processing abilities and spectro-temporal resolution. Although the original BEAR test battery also involved tests related to audibility and middle-ear analysis (as well as some additional tests in the four covered domains), the five tests were found to be the most informative and reliable for the classification of the listeners in auditory profiles.

The ACALOS test is able to estimate hearing thresholds, which are comparable to the ones provided by pure-tone audiometry (14). In the present study, these estimates were used as the most informative predictors for the fitted classifier (DT10) instead of the audiometric thresholds. ACALOS is also able to provide supra-threshold information related to loudness perception, such as the slope of the growth of loudness, the most comfortable level and the overall dynamic range. Therefore, the use of ACALOS could be of interest not only for the purpose of auditory profiling but also for hearing-aid fitting. For example, it would provide information about the growth of loudness of the patient that could guide fine-tuning of the gain at different input levels. Moreover, fitting formulas based on loudness normalization could be refined if loudness is measured with this technique (15).

The results of the fSTM test showed that profile C listeners have significantly poorer performance than listeners belonging to profiles A, B or D. This makes this new test quite interesting for classification. Additionally, the HINT results showed that profile B and C listeners had elevated speech reception thresholds in noise, suggesting that hearing-aid outcome will improve for these listeners if advanced processing is able to increase effectively improve the signal-to-noise ratio. Therefore, it would be interesting to investigate whether the tests involved in DT4 only (ACALOS, fSTM) are sufficient for a short version of the clinical test battery. Moreover, these tests are not language-dependent, in contrast to HINT. Although DT4 could be more easily adopted by the public health centers due to the cost-efficiency and shorter duration of the tests (35 min), the use of more informative tests, including the complete decision tree (DT10), should be of higher priority for a field study with research as the main purpose.

The unclassified group can only be identified using DT10 or DT7. It is of interest to identify this group during the field study further explore their supra-threshold auditory performance, which could help better understand the consequences of hearing loss in those listeners.

5 Conclusion

The results of the present study support the implementation of new audiological tests in the clinic to achieve a more comprehensive definition of the hearing abilities of patients with hearing loss. Four decision trees were evaluated in terms of classification performance and cost efficiency. The most informative and reliable tests beyond the audiogram were found to include the evaluation of spectro-temporal modulation sensitivity, loudness perception and binaural processing abilities. The BEAR clinical test battery will be evaluated in a large-scale study together with the new profile-based hearing-aid fitting strategies. The BEAR clinical test battery based on DT10, and proposed for such a field study, is available in a public repository[1].

ACKNOWLEDGEMENTS

We want to thank Dorte Hammershøi, James Harte and the other BEAR partners for their input during the realization of this study. This work was supported by Innovation Fund Denmark Grand Solutions 5164-00011B (Better hEAring Rehabilitation project), Oticon A/S, GN Hearing, Widex A/S, Aalborg University, University of Southern Denmark, the Technical University of Denmark, Force Technology, and the university hospitals in Aalborg, Odense and Copenhagen. The funding and input

[1] https://bitbucket.org/hea-dtu/bear-test-battery/

from all partners is sincerely acknowledged.

REFERENCES

1. Kochkin S. MarkeTrak VIII : Consumer satisfaction. Hear J. 2010;63(1):19.
2. Tecca J. Are post-fitting follow-up visits not hearing aid best practices? Hear Rev . 2018;25(4)(4):12–22. Available from: http://www.hearingreview.com/2018/03/post-fitting-follow-visits-not-hearing-aid-best-practices/
3. Sanchez Lopez R, Bianchi F, Fereczkowski M, Santurette S, Dau T. Data-Driven Approach for Auditory Profiling and Characterization of Individual Hearing Loss. Trends Hear . 2018 Jan;22:233121651880740. Available from: http://journals.sagepub.com/doi/10.1177/2331216518807400
4. Sanchez-Lopez R, Nielsen SG, El-Haj-Ali M, Bianchi F, Fereckzowski M, Cañete O, et al. Auditory tests for characterizing hearing deficits: The BEAR test battery. Int J Audiol. (in preparation).
5. Sanchez-Lopez R, Fereczkowski M, Neher T, Santurette S, Dau T. Robust auditory profiling: Improved data-driven method and profile definitions for better hearing rehabilitation. Proc ISAAR vol 7. 2019;(submitted).
6. Sweets JA, Pickett RM. Evaluation of Diagnostic Systems . Elsevier; 1982.
7. Gorga MP, Neely ST. Cost-effectiveness and test-performance factors in relation to universal newborn hearing screening. Mental Retardation and Developmental Disabilities Research Reviews. 2003.
8. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning . The Mathematical Intelligencer. New York, NY: Springer New York; 2009. 83–85 p. (Springer Series in Statistics; vol. 27).
9. Uus K, Bamford J, Taylor R. An analysis of the costs of implementing the National Newborn Hearing Screening Programme in England. J Med Screen . 2006 Mar 23;13(1):14–9. Available from: http://journals.sagepub.com/doi/10.1258/096914106776179764
10. Fleming S, Docs S. Commissioning Services for People with Hearing Loss: A Framework for Clinical Commissioning Groups . 2016 [cited 2019 May 14]. Available from: https://www.england.nhs.uk/publication/commissioning-hearing-loss-framework/
11. Mclean A. Development and Implementation of a National Funding and Service System for Hearing Aids Stage One Report March 2008. 2008;
12. Abrams H, Chisolm TH, McArdle R. A cost-utility analysis of adult group audiologic rehabilitation: are the benefits worth the cost? J Rehabil Res Dev . 39(5):549–58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17684833
13. Tecca J. Are post-fitting follow-up visits not hearing aid best practices? Hear Rev. 2018;25(4)(4):12–22.
14. Al-Salim SC, Kopun JG, Neely ST, Jesteadt W, Stiegemann B, Gorga MP. Reliability of Categorical Loudness Scaling and Its Relation to Threshold. Ear Hear . 2010 Aug [cited 2017 Jan 31];31(4):567–78. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20588122
15. Brand T, Hohmann V. An adaptive procedure for categorical loudness scaling. J Acoust Soc Am . 2002 Oct;112(4):1597–604. Available from: http://asa.scitation.org/doi/10.1121/1.1502902