Neural Prediction of Spoken Language Improvements in Children with Cochlear Implants

Yanlin Wang[1#], Di Yuan[1#], Shani Dettman[2], Dawn Choo[2], Emily Shimeng Xu[3], Denise Thomas[4], Maura E Ryan[5], Patrick C M Wong[1*], Nancy M Young[3,6,7*]

[1]Brain and Mind institute, The Chinese University of Hong Kong, Hong Kong SAR, China

[2]Department of Audiology & Speech Pathology, The University of Melbourne, 550 Swanston St, Parkville, Victoria 3010 Australia

[3]Division of Otolaryngology, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, Illinois, United States.

[4]Department of Audiology, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, Illinois, United States.

[5]Department of Medical Imaging, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, Illinois, United States.

[6]Department of Otolaryngology Head & Neck Surgery, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, United States.

[7]Knowles Hearing Center, Department of Communication Sciences and Disorders, Northwestern University, Evanston, Illinois, United States.

# These authors contributed equally to this work.

*Corresponding Authors:

Nancy M Young, MD

Division of Otolaryngology, Ann & Robert H. Lurie Children's Hospital of Chicago

Department of Otolaryngology Head & Neck Surgery, Feinberg School of Medicine, Northwestern University

Knowles Hearing Center, Department of Communication Sciences and Disorders, Northwestern University

Email: NYoung@luriechildrens.org

Patrick C M Wong, PhD

Brain and Mind institute, The Chinese University of Hong Kong

Email: p.wong@cuhk.edu.hk

Word Count: 3433

Abstract

Objective This study aims to construct neural predictive models to forecast post-CI spoken language improvements in children with hearing loss and to evaluate whether these models are language- and center-specific.

Methods A total of 278 children with hearing loss underwent magnetic resonance image (MRI) examinations and completed speech and language assessments both before and after the implants. We utilized deep transfer learning algorithms with pre-CI neuroanatomical features to predict post-CI spoken language development in children enrolled from 2009 to 2022, with 3-year follow-up.

Results We found that pre-CI MRI brain data can forecast spoken language development up to 36 months post-CI. Evidence of within-center and within-language prediction was consistent across different centers. MobileNet model exhibited the best performance with an accuracy (ACC) of 89.74% (95% CI, 89.39%-90.10%), sensitivity of 87.09% (95% CI, 86.17%-88.00%), specificity of 92.20% (95% CI, 90.98%-93.42%), and the area under the receiver operating characteristic curve (AUC) of 0.896 (95% CI, 0.893-0.900). However, cross-dataset generalization, even within the same center, could not be achieved with our current sample (e.g., ACC: 50.27% (95% CI, 47.62%-53.76%), sensitivity: 36.89% (95% CI, 0%-93.88%), specificity: 63.95% (95% CI, 6.43%-100%), and AUC: 0.499 (95% CI, 0.467-0.532)). When all the datasets were combined, the predictive performance remained high (ACC: 87.94% (95% CI, 87.28%-88.59%), sensitivity: 88.33% (95% CI, 97.18%-89.48%), specificity: 87.56% (95% CI, 86.12%-89.00%), and AUC: 0.879 (95% CI, 0.873-0.886)).

Conclusions The generalization of the neural predictive model across different centers and languages appears to be feasible and effective with a larger and more representative dataset.

Introduction

Cochlear implants (CI) have been shown to be effective in assisting children with severe to profound hearing loss to develop spoken language.[1] However, many children with CI still lag behind their peers with normal hearing in terms of spoken language development.[2-4] Despite the availability of various early intervention approaches such as listening and spoken language therapy with or without sign language, there is little consensus on the optimal type and dose of intervention.[5] Accurately predicting spoken language development on the individual child level prior to CI would allow for the provision of more intensive healthcare for those children who may need it most.

It has been demonstrated that brain measures often serve as better prognostic indicators, either alone or in combination with other measures, than traditional measures such as age at implant and pre-implantation residual hearing.[6] Studies have successfully used machine learning techniques to forecast the auditory and spoken language skills of children with CI.[7,8] For example, the preoperative neuroanatomical features of CI users predicted the variability of their speech perception improvements six months after surgery, showing 84% accuracy based on a linear support vector machine (SVM) classifier with a recursive feature elimination selection technique.[8] In contrast, non-neural features, including demographic variables and pre-CI speech perception scores only reached a chance level of accuracy in predicting speech perception improvements. The robustness and efficiency of brain measures in predicting post-CI improvements have also been supported by studies using preoperative brain activations in response to audio and visual stimuli in children and adults with CI.[9,10]

It is worth noting that the correlation between preoperative brain measures and post-CI outcomes cannot provide sufficient prognostic values at an individual level, although the findings may illustrate the neural basis of spoken language development in people with CI.[11–13] Moreover, predicting the improvements from pre- to post-CI might be more important than predicting post-CI outcomes. This is because the outcomes measured after implantation are usually closely correlated with pre-implantation measures,[8,14] and the correlation between brain measures and post-CI outcomes could be confounded by the baseline measures. Children with poor speech abilities before implantation may still demonstrate significant improvements due to the benefits provided by CI. As supported in a previous study, children’s pre-CI speech perception ability was independent of their improvements after receiving CI.[8] Therefore, predicting the change in spoken language of pediatric CI users provides more information related to CI benefits. This allows for guiding precision healthcare, enabling timely adjustments to intervention plans, and helping manage parental expectations of children's post-CI improvements. Ultimately, accurate prediction on the individual child level enabled by our approach will permit the optimization of spoken language and an improved quality of life after CI.

Although a predictive model utilizing preoperative brain measures has been built by our research group to forecast improvements of the spoken language measures, training of the predictive models were restricted to children from a single medical center and to children learning English. For both clinical and theoretical reasons, it is important to ascertain whether neural predictive models constructed with data from one medical center and one language can be used to predict the improvements of children who are from other medical centers and learning other languages. From the clinical standpoint, model generalization means that it is unnecessary to construct population-specific predictive models, as reliance on models constructed with data from a variety of patients from any center would be sufficient. From the theoretical standpoint, generalization speaks to the basic neural architecture subserving language development. Do the networks that support English learning substantially overlap with those supporting Spanish or Cantonese learning?

Our multicenter study aimed to address the question of model generalization with a deep-learning model predicting children's spoken language improvements up to three years after implantation. Because of the low rate of severe to profound hearing loss in children, it is unusual to have a dataset large enough to train a predictive model. This study employed a transfer learning architecture, leveraging the learned features from pre-trained models on large-scale image datasets to enhance the performance of our own model.[15]

Methods

Participants

Children with congenital or early onset sensorineural hearing loss were recruited from three different centers: Chicago, United States; Melbourne, Australia; and Hong Kong, China. They received CI at local hospitals from 2009 to 2022. All the children underwent T1-weighted structural whole-brain magnetic resonance imaging (MRI) as a part of their pre-CI evaluation. Their speech and language abilities were assessed before and after implantation. Parents or guardians provided written informed consent to access children's MRI scans and clinical data. This study was approved by the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee, the Stanley Manne Children's Research Institute's Institutional Review Board, and The Royal Children's Hospital, Human Research Ethics Committee at each center.

As a study aiming to predict improvements in as many children with CI as possible, we imposed relatively broad inclusion/exclusion criteria. At each center, children had to be from homes that speak Cantonese (Hong Kong), English (Melbourne), or English or Spanish (Chicago) as the dominant language. We excluded children who had a known genetic condition that is expected to severely affect language development and children who had gross brain malformations. A total of 278 children were included. The demographic information is shown in Table 1.

Clinical Measures

Children's auditory skill, speech perception, receptive and/or expressive language abilities were measured before and up to 36 months after implantation using different assessment tools across centers (see the Supplementary Materials). We here refer to all these measurements as ‘spoken language,’ being aware that audition and speech perception are precursors for spoken language development.[16,17] Positive correlations have been demonstrated between speech perception and spoken language scores on standardized tests for children with hearing loss.[18,19] While variances could be introduced by differences in the assessment methods and timing, it is feasible to compare the spoken language ability across the centers and over time because of the heterotypic stability inherent in spoken language development.[20,21] Specifically, the individual ranking of different manifest characteristics is maintained over time as long as those characteristics share the same underlying construct and theoretical value.

The improvement of spoken language development from pre- to post-CI was quantified by the change of assessed scores as a function of assessment time for each participant. To this end, a linear mixed-effect model was constructed for each center with spoken language scores as the dependent variable, subject ID as a random intercept, as well as assessment time as a random slope. The fixed effects portion of the model included only the intercept term, as the influence of time on spoken language scores was captured in the random slope. The model can be expressed mathematically as Scores ~ 1 + (assessment time | subject ID). The random slope in the model allowed us to estimate individual differences in the rate of speech and language change over time. For better model generalization, instead of using the raw scores directly for fine-grained prediction, we separated the spoken language improvement into binary classifications (high-improvement and low-improvement) using a median split approach within each center.

MRI acquisition and preprocessing

The T1-weighted MRI image was obtained from each child before CI. The scanning parameters were optimized to obtain a good signal-to-noise ratio (Supplementary Material). MRI images were processed using the Advanced Normalization Tools (ANTs) in Python.[22] To increase the image quality, the images were resampled to 1 mm× 1 mm× 1 mm voxel size and preprocessed following the basic preprocessing pipeline for T1-weighted brain MRI in ANTs. The deformation-based morphometry (DBM) method was used to examine the morphological differences over the entire brain with an age appropriate T1 image as the template.[23,24] Fifteen axial 2D slices were extracted from the central part of the 3D DBM brain scans.[25] The images were cropped and resized into a target resolution of 128×128 voxels and were normalized using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) before being passed on for further analyses.[26] Each slice was assigned the same label as the corresponding subject and used as a data sample to train the model.

Transfer Learning and Feature Extractions

We utilized popular pre-trained convolutional neural network (CNN) models, including AlexNet,[27] VGG19,[28] ResNet,[29] Inception,[30] GoogleNet,[31] MobileNet,[32] and DenseNet,[33] implemented in PyTorch version 1.9, for feature extraction. This standard transfer learning strategy involves using pre-trained CNN models on ImageNet as the backbone of the model to capture generic and domain-specific features, followed by fine-tuning the top layers to learn new specialized representations tailored to our output classifier.[26,34] During the fine-tuning phase, the weights and biases of the CNN models were frozen to prevent changes. Due to differences in the CNN architectural designs, an adaptive pooling operation was applied to AlexNet and MobileNet before the final classification layer to ensure that the output became a one-dimensional vector. Subsequently, a new fully connected layer, the classification layer, was added to process the outputs from the hidden layer's activation function and compose the final classification. Data augmentation with random rotation and flipping was executed to improve the model training efficiency.[35,36] The loss function was binary cross-entropy with logit loss. The optimizer was Adam with a learning rate of 1×10-4. A total of 200 epochs with a batch size of 64 images were set for training. The validation performance was used to determine when to stop the training. The CNN models were trained until there was no improvement in the validation loss for 10 consecutive epochs. All the experiments were conducted by dividing the data into 80% for training and validation and 20% for held-out testing. A five-fold cross-validation approach was used to validate the model's performance during training. The training validation results were obtained through this five-fold cross-validation process to detect language improvements. Finally, a held-out 20% test set was used to evaluate the model's performance, specifically its generalization.

Performance comparisons

To examine whether neural features can predict longer-term post-CI improvements, we first compared state-of-the-art CNN models within a single center (Chicago or Melbourne) or a single language dataset (English or Spanish). To further assess the generalization of the predictive model with a new dataset that has different inclusion criteria or was obtained from different facilities, we tested whether model trained on the largest (Chicago English) dataset could predict improvements for CI candidates learning Spanish at the same center, or for CI candidates learning the same language at another medical center. These external assessments across different languages or centers were conducted on trained model on a single dataset. Finally, to assess the robustness and generalization of the predictive model on combined dataset, we developed the model on a development set (80%) of the combined dataset across centers and languages, which was then internally validated using an held-out 20% test dataset from the same combined dataset. In addition, we also compared sliced-based CNN models with voxel-based machine learning models including Linear regression (LR), SVM, Random Forest (RF), Decision Tree (DT), K-

Nearest Neighbor (KNN), and eXtreme Gradient Boosting (XGBoost) (see Supplementary Materials).

Performance Evaluation Metrics

The model's performance in classification could be evaluated using the following performance metrics: the area under the receiver operating characteristic curve (AUC), accuracy (ACC), sensitivity, and specificity. AUC measures the model's ability to discriminate between classes across various thresholds and is calculated from the False Positive Rate (FPR) and True Positive Rate (TPR). ACC measures the proportion of correctly classified images, reflecting the overall effectiveness of the model. Sensitivity, or recall, assesses the classifier's ability to correctly identify cases with the disease. Specificity evaluates how well the classifier can identify cases without the disease.

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN}) / (\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$$

$$\mathrm{Sensitivity} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN})$$

$$\mathrm{Specificity} = \mathrm{TN} / (\mathrm{FP} + \mathrm{TN})$$

where TP is true positive values, TN is true negative values, FP is false positive values, and FN is false negative values; is a positive instance and is a negative instance.

Results

Children with CI showed improvements in spoken language abilities compared to the baseline measurement tested before implantation (Figure 1). Specifically, in Chicago, the spoken language abilities of English-learning children improved from 75 to 292, and of Spanish-learning children from 45 to 203, over the period from pre-CI to 36 months post-CI, as tested by SRI-m. Similarly, in Hong Kong, Cantonese-learning children improved from 17 to 32, over the period from pre-CI to 24 months post-CI, as tested by LittlEARS. Most of these improvements emerged in the first year and a half after implantation. In Melbourne, the receptive language of English-learning children improved from 74 to 85 in the first two years after implantation but dropped to 70 in the third year post-CI, as tested by PPVT and PLS. The different pattern of changes in spoken language development may result from the standard scores obtained in Melbourne, which take age-appropriate normal-hearing children as a control, suggesting that children were able to catch up with their normal-hearing peers but still lagged behind in their long-term spoken language development. Despite different standardized tests being used to capture the spoken language development across the centers, our predictive models were constructed to only predict the binary classifications of low or high improvement.

Table 2 lists the deep learning and machine learning models' training and testing ACC, sensitivities, specificities, and AUC. In general, slice-based deep transfer learning can substantially improve the model’s prediction performance compared to voxel-based machine learning models on Chicago English data (Figure 2A). Among the various deep learning convolutional neural network models, the MobileNet model exhibits the best performance with an ACC of 89.74% (95% CI, 89.39%-90.10%), sensitivity of 87.09% (95% CI, 86.17%-88.00%), specificity of 92.20% (95% CI, 90.98%-93.42%), and AUC of 0.896 (95% CI, 0.893-0.900) on the test dataset. Predictive models using slice-based deep transfer learning can achieve a high level of predictive performance when a single dataset is used (e.g., data from English-learning children from Chicago were used to test the same model). Therefore, we used the MobileNet model as a baseline network for downstream assessments of model's generalization.

However, when the generalization was externally tested using data from another medical center (e.g., testing the Chicago English model with Melbourne English data), the model's performance dropped to chance levels (ACC: 50.95% (95% CI, 49.14%-53.75%), sensitivity: 62.90% (95% CI, 3.74%-100%), specificity: 39.28% (95% CI, 0%-95.66%), and AUC: 0.511 (95% CI, 0.489-0.533)) (Table 3 and Figure 2B). Even within the same center, cross-language generalization (e.g., testing the Chicago English model with Chicago Spanish data) could not be achieved with our sample sizes (ACC: 50.27% (95% CI, 47.62%-53.76%), sensitivity: 36.89% (95% CI, 0%-93.88%), specificity: 63.95% (95% CI, 6.43%-100%), and AUC: 0.499 (95% CI, 0.467-0.532)). When tested across different languages and cultural backgrounds (e.g., testing the Chicago English model with Hong Kong Cantonese data), the model showed an ACC of 50.75% (95% CI, 47.62%-53.87%), sensitivity of

36.67% (95% CI, 0%-96.18%), specificity of 63.26% (95% CI, 3.46%-100%), and AUC of 0.500 (95% CI, 0.496-0.504).

Nevertheless, regardless of whether a single dataset or a combination of different datasets was used to build the model, the MobileNet model demonstrated consistently accurate performance (Table 3 and Fig 2B). It achieved an ACC of 87.38% (95% CI, 87.12%-87.64%), sensitivity of 85.36% (95% CI, 84.02%-86.70%), specificity of 89.57% (95% CI, 88.04%-91.11%), and AUC of 0.874 (95% CI, 0.871-0.876) across the Chicago and Melbourne datasets. When tested across the Chicago, Melbourne, and Hong Kong datasets, it achieved an ACC of 87.94% (95% CI, 87.28%-88.59%), sensitivity of 88.33% (95% CI, 87.18%-89.48%), specificity of 87.56% (95% CI, 86.12%-89.00%), and AUC of 0.879 (95% CI, 0.873-0.886).

Discussion

In this multicenter study, we employed the transfer deep learning technique using the preoperative neuroanatomical features to forecast spoken language improvements in children with CI for up to three years. Our transfer learning models consistently demonstrated accurate performance in distinguishing between higher and lower improvement groups for both single dataset and combined datasets. However, the models exhibited poor performance when applied to external generalization testing. The findings highlight the effectiveness of using transfer deep learning to predict post-CI improvements on the individual-child level for the precision care of pediatric CI users. The poor generalization in external testing, however, calls for multicenter collaboration to obtain a large-scale representative data, enabling the construction of models with better potential to generalize to new patients from diverse backgrounds.

Transfer learning offers an effective strategy for the target domain classifier by integrating the knowledge learned from pre-trained CNN models on ImageNet with new specialized representations through fine-tuning.[15,34] This approach has shown to be powerful in healthcare decisions for rare diseases, such as Alzheimer’s disease,[37] cardiomyopathy,[38] diabetic retinopathy,[39] etc. Compared to a previous study that used voxel-based machine learning models (i.e., SVM) to predict speech perception improvements six months post-CI with 37 children,[8] our study employing a transfer learning approach revealed a higher prediction accuracy even for longer-term post-CI improvements using a larger sample size. Our study is among the first to use such a transfer learning approach for predicting children's post-CI improvements

Generalization to a new dataset is crucial for ensuring the applicability and real-world impact of any scientific findings.[40,41] A universal model is desirable for generalizing across datasets. In this study, model trained on a single dataset were unable to generalize directly to other datasets with different cultural or language characteristics. Although the model achieved a test accuracy of 89.74% (95% CI, 89.33%-90.10%) for the Chicago English dataset, external generalization testing on new datasets resulted in poor predictive performance of 50.95% (95% CI, 49.11%-52.75%) accuracy for the Melbourne English dataset, 50.27% (95% CI, 46.78%-53.76%) for the Chicago Spanish dataset, and 50.75% (95% CI, 47.62%-53.87%) for the Hong Kong Cantonese dataset. These independent datasets shared the same language but had different cultural backgrounds (Melbourne English), shared the same cultural background but had different language experiences (Chicago Spanish), or had completely different language and cultural backgrounds (Hong Kong Cantonese). The poor generalization of the model may result from the heterogeneous languages and cultural backgrounds across the datasets making the unseen data mismatch the training distribution. It has been demonstrated that cultural and language differences have a large impact on brain function and structure.[42,43] Thus, generalization across datasets will require the incorporation of subjects from diverse cultural and language backgrounds, allowing the model to learn additional features during training to avoid characteristic-specific model and enable robust generalization.

Furthermore, we investigated the generalization of the predictive model on the combined dataset. Accordingly, the model was trained on the development set (80%) of the combined dataset across centers and languages, and internal validation was conducted on a held-out 20% of the test dataset. The performance of these models trained on combined datasets showed consistently higher accuracies compared to those trained on a single dataset. These findings demonstrated that the preoperative neural features can significantly predict post-CI improvements in children with hearing loss from different languages and centers. Moreover, the transfer learning strategy can effectively adapted to combined datasets with different cultural or language characteristics, enhancing the robustness and generalization of model. Our results imply that, ultimately, it is possible to improve the generalization across different populations using transfer learning techniques and more representative datasets, which is critical for the future translation to clinical practice.

Our study had several limitations. First, although the study included diverse participants with datasets from multiple centers and languages, the sample size was relatively small, which might not be sufficiently diversity for developing a universal model as a pre-surgical screening tool. Second, different assessment tools were used across centers. While it would be ideal to use unified tools for better generalization, we conducted the binary classifications (high improvement and low improvement) using a median split approach. This accommodates the measurements taken on different scales across the centers. Third, the limitation of spatial information between slices, as each 2D slice is processed independently,[44] was mitigated by using transfer learning and fine-tuning techniques to integrate prior knowledge from large datasets with domain-specific knowledge. Future research should focus on testing the model's generalization across diverse populations and settings, including CI children from different centers and cultural backgrounds.

Conclusions

Our study demonstrated that the deep transfer learning approach provides an effective means for utilizing preoperative brain images to predict whether children will have high or low spoken language improvements after CI. Furthermore, assessments of the model's generalization demonstrated that while model trained on a single dataset cannot directly generalize to a new dataset with different cultural or language characteristics, those trained on combined datasets showed better performance, highlighting the need for multicenter collaboration to generate a large, diverse dataset for the purpose of building a universal model to forecast spoken language development in children with CI.

References

1. Sharma SD, Cushing SL, Papsin BC, Gordon KA. "Hearing and speech benefits of cochlear implantation in children: A review of the literature." *International Journal of Pediatric Otorhinolaryngology*. 2020;133:109984. doi:10.1016/j.ijporl.2020.109984.

2. Ching TYC, Dillon H, Button L, et al. "Age at Intervention for Permanent Hearing Loss and 5-Year Language Outcomes." *Pediatrics*. 2017;140(3):e20164274. doi:10.1542/peds.2016-4274.

3. Ching TYC, Dillon H, Marnane V, et al. "Outcomes of Early- and Late-identified Children at 3 Years of Age: Findings from a Prospective Population-based Study." *Ear Hear*. 2013;34(5):535-552. doi:10.1097/AUD.0b013e3182857718.

4. Karltorp E, Eklöf M, Östlund E, Asp F, Tideholm B, Löfkvist U. "Cochlear implants before 9 months of age led to more natural spoken language development without increased surgical risks." *Acta Paediatrica*. 2020;109(2):332-341. doi:10.1111/apa.14954.

5. Chu C, Dettman S, Choo D. "Early intervention intensity and language outcomes for children using cochlear implants." *Deafness & Education International*. 2020;22(2):156-174. doi:10.1080/14643154.2019.1685755.

6. Gabrieli JD, Ghosh SS, Whitfield-Gabrieli S. "Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience." *Neuron*. 2015;85(1):11-26.

7. Yuan D, Chang WT, Ng IHY, et al. "Preoperative Neuroanatomical Features Outperform Non-Neural Features in Predicting Auditory Skills in Chinese-Learning Children After Cochlear Implantation." Published online May 31, 2024. doi:10.31234/osf.io/e2w5y.

8. Feng G, Ingvalson EM, Grieco-Calub TM, et al. "Neural preservation underlies speech improvement from auditory deprivation in young cochlear implant recipients." *PNAS*. 2018;115(5):E1022-E1031. doi:10.1073/pnas.1717603115.

9. Tan L, Holland SK, Deshpande AK, Chen Y, Choo DI, Lu LJ. "A semi-supervised Support Vector Machine model for predicting the language outcomes following cochlear implantation based on pre-implant brain fMRI imaging." *Brain and Behavior*. 2015;5(12):e00391. doi:10.1002/brb3.391.

10. Kyong JS, Suh MW, Joon Han J, et al. "Cross-Modal Cortical Activity in the Brain Can Predict Cochlear Implantation Outcome in Adults: A Machine Learning Study." *J Int Adv Otol*. 2021;17(5):380-386. doi:10.5152/iao.2021.9337.

11. Giraud AL, Lee HJ. "Predicting cochlear implant outcome from brain organisation in the deaf." *Restorative neurology and neuroscience*. 2007;25(3-4):381-390.

12. Anderson CA, Wiggins IM, Kitterick PT, Hartley DEH. "Adaptive benefit of cross-modal plasticity following cochlear implantation in deaf adults."

Proceedings of the National Academy of Sciences. 2017;114(38):10256-10261. doi:10.1073/pnas.1704785114.

13. Lee HJ, Giraud AL, Kang E, et al. "Cortical activity at rest predicts cochlear implantation outcome." *Cerebral Cortex*. 2007;17(4):909-917.

14. Kim H, Kang WS, Park HJ, et al. "Cochlear Implantation in Postlingually Deaf Adults is Time-sensitive Towards Positive Outcome: Prediction using Advanced Machine Learning Techniques." *Sci Rep*. 2018;8(1):18004. doi:10.1038/s41598-018-36404-1.

15. Iman M, Arabnia HR, Rasheed K. "A review of deep transfer learning and recent advancements." *Technologies*. 2023;11(2):40.

16. Perigoe CB, Paterson MM. "Understanding auditory development and the child with hearing loss." *Fundamentals of audiology for the speech-language pathologist*. Published online 2013:173-204.

17. Werker JF, Hensch TK. "Critical Periods in Speech Perception: New Directions." *Annu Rev Psychol*. 2015;66(1):173-196. doi:10.1146/annurev-psych-010814-015104.

18. Geers AE, Nicholas JG, Sedey AL. "Language skills of children with early cochlear implantation." *Ear and hearing*. 2003;24(1):46S-58S.

19. DesJardin JL, Ambrose SE, Martinez AS, Eisenberg LS. "Relationships between speech perception abilities and spoken language skills in young children with hearing loss." *International Journal of Audiology*. 2009;48(5):248-259. doi:10.1080/14992020802607423.

20. Bornstein MH, Putnick DL, Esposito G. "Continuity and Stability in Development." *Child Development Perspectives*. 2017;11(2):113-119. doi:10.1111/cdep.12221.

21. Bornstein MH, Hahn CS, Putnick DL, Pearson RM. "Stability of core language skill from infancy to adolescence in typical and atypical development." *Sci Adv*. 2018;4(11):eaat7422. doi:10.1126/sciadv.aat7422.

22. Tustison NJ, Cook PA, Holbrook AJ, et al. "The ANTsX ecosystem for quantitative biological and medical imaging." *Sci Rep*. 2021;11(1):9068. doi:10.1038/s41598-021-87564-6.

23. Gaser C, Nenadic I, Buchsbaum BR, Hazlett EA, Buchsbaum MS. "Deformation-based morphometry and its relation to conventional volumetry of brain lateral ventricles in MRI." *Neuroimage*. 2001;13(6 Pt 1):1140-1145. doi:10.1006/nimg.2001.0771.

24. Shi F, Yap PT, Wu G, et al. "Infant Brain Atlases from Neonates to 1- and 2-Year-Olds." *PLOS ONE*. 2011;6(4):e18746. doi:10.1371/journal.pone.0018746.

25. Wen J, Thibeau-Sutre E, Diaz-Melo M, et al. "Convolutional neural networks for classification of Alzheimer’s disease: Overview and reproducible evaluation." *Medical image analysis*. 2020;63:101694.

26. Ardalan Z, Subbian V. "Transfer learning approaches for neuroimaging analysis: a scoping review." *Frontiers in artificial intelligence*. 2022;5:780405.

27. Krizhevsky A, "Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012;25. Accessed June 7, 2024. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

28. Simonyan K, Zisserman A. "Very Deep Convolutional Networks for Large-Scale Image Recognition." Published online April 10, 2015. Accessed June 7, 2024. http://arxiv.org/abs/1409.1556.

29. He K, Zhang X, Ren S, Sun J. "Deep residual learning for image recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016:770-778. Accessed June 7, 2024. http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

30. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. "Rethinking the inception architecture for computer vision." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016:2818-2826. Accessed June 7, 2024. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.

31. Szegedy C, Liu W, Jia Y, et al. "Going deeper with convolutions." In *Proceedings of the ieee conference on computer vision and pattern recognition*; 2015. *Google Scholar*. Published online 2015:1-9.

32. Howard AG, Zhu M, Chen B, et al. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." Published online April 16, 2017. Accessed June 7, 2024. http://arxiv.org/abs/1704.04861.

33. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. "Densely connected convolutional networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2017:4700-4708. Accessed June 7, 2024. http://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html.

34. Yosinski J, Clune J, Bengio Y, Lipson H. "How transferable are features in deep neural networks?" *Advances in neural information processing systems*. 2014;27. Accessed June 7, 2024. https://proceedings.neurips.cc/paper_files/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html.

35. Taqi AM, Awad A, Al-Azzo F, Milanova M. "The impact of multi-optimizers and data augmentation on TensorFlow convolutional neural network performance." In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE; 2018:140-145. Accessed June 7, 2024. https://ieeexplore.ieee.org/abstract/document/8396988/.

36. Afzal S, Maqsood M, Nazir F, et al. "A data augmentation-based framework to handle class imbalance problem for Alzheimer’s stage detection." *IEEE access*. 2019;7:115528-115539.

37. Saleh AW, Gupta G, Khan SB, Alkhaldi NA, Verma A. "An Alzheimer’s disease classification model using transfer learning Densenet with embedded healthcare decision support system." *Decision Analytics Journal*. 2023;9:100348. doi:10.1016/j.dajour.2023.100348.

38. Theodoris CV, Xiao L, Chopra A, et al. "Transfer learning enables predictions in network biology." *Nature*. 2023;618(7965):616-624. doi:10.1038/s41586-023-06139-9.

39. Dai L, Wu L, Li H, et al. "A deep learning system for detecting diabetic retinopathy across the disease spectrum." *Nat Commun*. 2021;12(1):3242. doi:10.1038/s41467-021-23458-5.

40. Alexander GC, Emerson S, Kesselheim AS. "Evaluation of aducanumab for Alzheimer disease: scientific evidence and regulatory review involving efficacy, safety, and futility." *Jama*. 2021;325(17):1717-1718.

41. Kriegeskorte N, Mur M, Bandettini PA. "Representational similarity analysis-connecting the branches of systems neuroscience." *Frontiers in systems neuroscience*. 2008;2:249.

42. Paulesu E, McCrory E, Fazio F, et al. "A cultural effect on brain function." *Nat Neurosci*. 2000;3(1):91-96. doi:10.1038/71163.

43. Han S, Northoff G. "Culture-sensitive neural substrates of human cognition: A transcultural neuroimaging approach." *Nature reviews neuroscience*. 2008;9(8):646-654.

44. Sarraf S, DeSouza DD, Anderson J, Tofighi G, Initiativ ADN. "DeepAD: Alzheimer’s disease classification via deep convolutional neural networks using MRI and fMRI." *BioRxiv*. Published online 2016:070441.

Figure Legends:

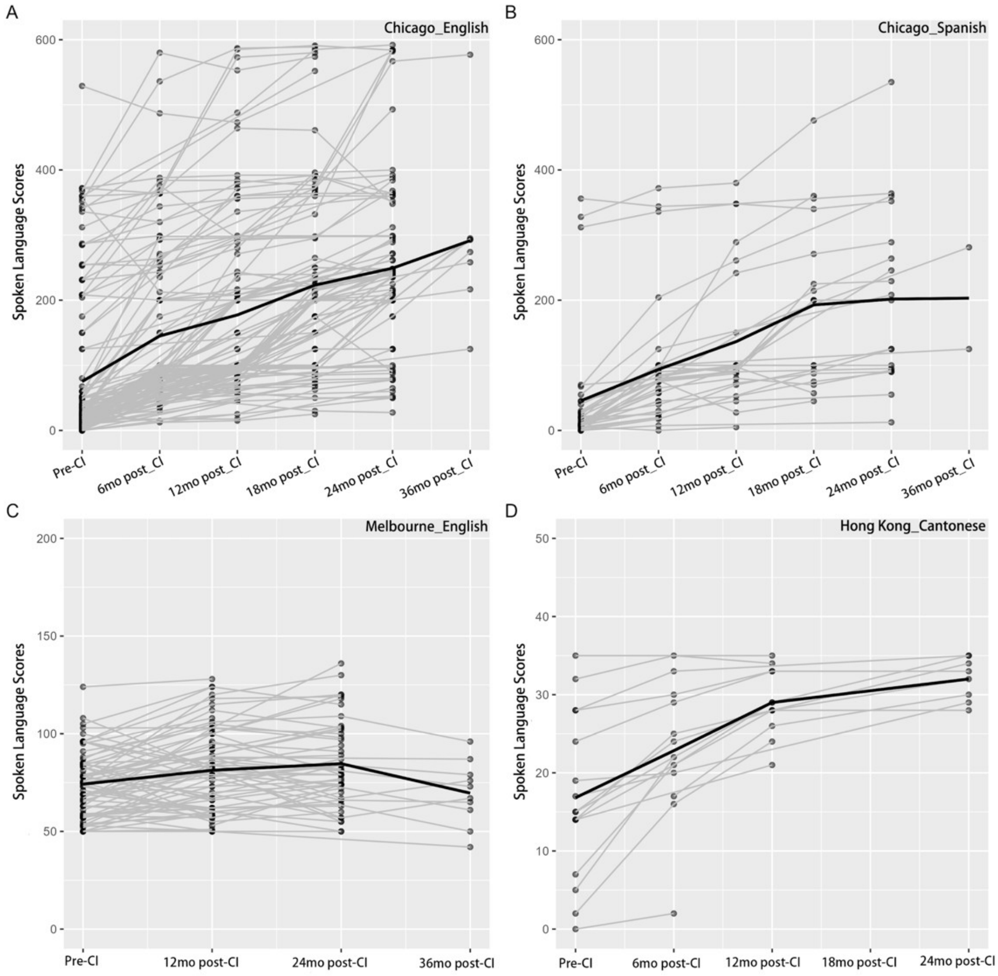

Figure1. Spoken language ability of children from before to after implantation at each center. The dots and gray lines indicate the change of spoken language scores for each individual across time. The black line indicates the mean change of spoken language scores for all children at

each center.

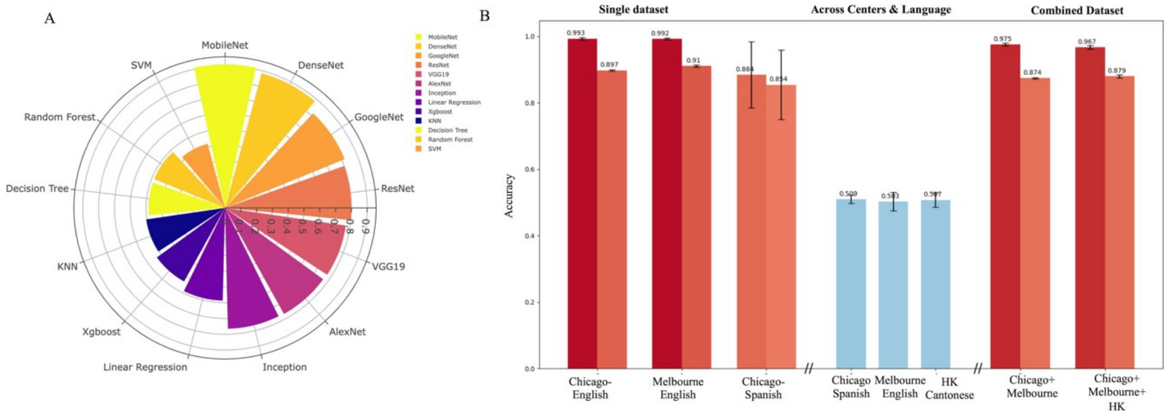

Figure2. Performance comparison for machine learning models and transfer learning models (A) and assessments of transfer learning model's generalization on single datasets, external test datasets, and combined datasets (B). Error bars represent plus/minus one standard deviation, showing the means of accuracy with standard deviations across five-fold cross-validation from different experiments.

Table 1. Demographic information for participants from different centers.

	Chicago data		**Melbourne data**	**Hong Kong data**	**All**
Sample size	143	37	81	17	278
Family language	English	Spanish	English	Cantonese	NA
Female, No. (%)	67 (46.9)	21 (56.8)	37 (45.7)	12 (70.6□	137 (49.3)
Age at SNHL diagnosis, mean (SD), mo	10.2 (13.3)	11.1 (12.4)	3.2 (4.4)	11.6 (15.2)	9.7 (12.8)
Age of HA fitting, mean (SD), mo	11.6 (13.2)	12.3 (12.5)	3.8 (4.2)	16.9 (13.6)	10.4 (12.3)
Age at MRI, mean (SD), mo	23.8 (20.5)	26.9 (18.2)	11.4 (12.1)	24.3 (18.0)	20.7 (18.9)
Age at CI, mean (SD), mo	27.4 (20.9)	30.1 (18.4)	19.2 (13.2)	32.5 (16.6)	25.7 (18.8)
Unaided hearing of left ear, dB HL	95.4 (17.0)	98.9 (18.0)	97.7 (18.7)	103.3 (15.7)	96.9 (17.5)
Unaided hearing of right ear, dB HL	93.7 (18.1)	100.2 (15.1)	99.5 (19.0)	101.7 (14.0)	96.5 (17.9)

Abbreviations: CI, cochlear implants; MRI, magnetic resonance imaging; HA, hearing aid; SNHL, sensorineural hearing loss; NA, not applicable

Table 2. The classification performance of the Transfer Learning models and Machine Learning models in the Chicago English group.

Types	Models	% (95% CI)			
		Accuracy	Sensitivity	Specificity	AUC (95% CI)
Slice-based	VGG19_bn	81.17 (80.11-82.22)	86.19 (84.80-87.57)	75.73 (73.55-77.90)	0.810 (0.799-0.820)
	ResNet-50d	88.02 (86.92-89.11)	88.16 (85.98-90.34)	87.86 (86.21-89.51)	0.880 (0.869-0.891)
	DenseNet_169	89.09 (88.06-90.12)	92.11 (91.47-92.74)	85.83 (83.64-88.02)	0.890 (0.879-0.900)
	AlexNet	79.95 (78.61-81.30)	84.13 (82.67-85.58)	75.44 (72.53-78.35)	0.800 (0.786-0.813)
	Inception_V3	83.64 (81.75-85.53)	85.65 (77.40-93.90)	81.46 (73.24-89.67)	0.836 (0.817-0.854)
	GoogleNet	87.13 (85.54-88.72)	92.38 (90.53-94.22)	81.46 (79.07-83.84)	0.869 (0.853-0.885)
	MobileNet	89.74 (89.39-90.10)	87.09 (86.17-88.00)	92.20 (90.98-93.42)	0.896 (0.893-0.900)
Voxel-based	LR	58.74 (47.71-69.77)	52.89 (41.88-63.91)	63.51 (31.67-95.34)	0.582 (0.432-0.732)
	DT	55.30 (37.53-73.07)	74.65 (53.49-95.81)	38.43 (9.40-67.46)	0.565 (0.477-0.654)
	SVM	49.73 (40.55-58.91)	36.67 (8.26-65.08)	63.40 (34.43-92.37)	0.500 (0.414-0.586)
	KNN	50.37 (43.68-57.06)	53.25 (28.96-77.55)	47.54 (22.54-72.54)	0.504 (0.431-0.577)
	RF	48.45 (31.79-65.11)	36.38 (15.65-57.12)	66.13 (35.02-97.25)	0.5123 (0.364-0.661)
	XGBoost	53.25 (42.39-64.12)	53.86 (42.30-65.43)	53.07 (34.47-71.66)	53.47 (41.35-65.58)

Abbreviations: LR, Logistic Regression; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; DT, Decision Tree; RT, Random Forest; XGBoost, eXtreme Gradient Boosting.

Table 3. The performance of the Transfer Learning method within and across datasets using the MobileNet model.

Datasets		% (95% CI)			
		Accuracy	Sensitivity	Specificity	AUC (95% CI)
Single Dataset	Chicago_English	89.74 (89.39-90.10)	87.09 (86.17-88.00)	92.20 (90.98-93.42)	0.896 (0.893-0.900)
	Melbourne_English	91.03 (90.60-91.46)	91.67 (90.63-92.70)	90.41 (89.09-91.72)	0.910 (0.906-0.915)
	Chicago_Spanish	85.41 (70.96-99.85)	89.02 (87.69-90.35)	82.33 (54.97-99.96)	0.857 (0.724-0.990)
Across Center	Melbounre_English[a]	50.95 (49.14-52.75)	62.90 (3.74-100)	39.28 (0-95.66)	0.511 (0.489-0.533)
Across Language	Chicago_Spanish[a]	50.27 (46.78-53.76)	36.89 (0-93.88)	63.95 (6.43-100)	0.499 (0.467-0.532)
Across Center & Language	Hong Kong_Cantonese[a]	50.75 (47.62-53.87)	36.67 (0-96.18)	63.26 (3.46-100)	0.500 (0.496-0.504)
Combined Dataset	Chicago+Melbourne	87.38 (87.12-87.64)	85.36 (84.02-86.70)	89.57 (88.04-91.11)	0.874 (0.871-0.876)
	Chicago+Melbourne+Hong Kong	87.94 (87.28-88.59)	88.33 (87.18-89.48)	87.56 (86.12-89.00)	0.879 (0.873-0.886)

[a] The external validation across different languages or centers was conducted on trained model on a single dataset (Chicago English) separately.