Does artificial intelligence have a role in telehealth screening of ear disease in Indigenous children in Australia?

Adam Mothershaw[1,2]^, Anthony C. Smith[1,2,3]^, Christopher F. Perry[4,5], Cecil Brown[6], Liam J. Caffery[1,2]^

[1]Centre for Online Health, The University of Queensland, Brisbane, Australia; [2]Centre for Health Services Research, The University of Queensland, Brisbane, Australia; [3]Centre for Innovative Medical Technology, University of Southern Denmark, Odense, Denmark; [4]Faculty of Medicine, The University of Queensland, Brisbane, Australia; [5]Paediatric Otolaryngology Head and Neck Surgery Department, Queensland Children's Hospital, Brisbane, Australia; [6]Cherbourg Community Health Service, Darling Downs Health, Cherbourg, Australia

Contributions: (I) Conception and design: A Mothershaw, LJ Caffery; (II) Administrative support: A Mothershaw, AC Smith; (III) Provision of study material or patients: All authors; (IV) Collection and assembly of data: A Mothershaw, CF Perry, C Brown; (V) Data analysis and interpretation: A Mothershaw, LJ Caffery; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Associate Professor Liam J. Caffery. Ground Floor, Building 33, Princess Alexandra Hopsital, Woolloongabba QLD 4102, Australia. Email: l.caffery@coh.uq.edu.au.

Background: Artificial intelligence models have been used to successfully detect ear disease from video-otoscopic images. Artificial intelligence is highly context specific, and to date there are no studies looking specifically at classifying ear disease amongst Aboriginal and Torres Strait Islander children. The aim of this study was to establish proof-of-concept of both the feasibility and effectiveness of using artificial intelligence to detect ear disease in a telehealth ear screening service for Indigenous children.

Methods: Labelled video-otoscopic images from the Health-e-Screen4kids service located in Cherbourg, Queensland were used to train an artificial intelligence binary classifier model of ear disease. The model uses video-otoscopic images as input and outputs a binary prediction of whether the ear is normal or abnormal. The accuracy of the resultant model was assessed by comparing the output of the model to the image labelled by an Ear, Nose, and Throat (ENT) specialist.

Results: The artificial intelligence model achieved an overall accuracy of 80.99% on the validation set (images used for training the model) and 78.90% on the test set (images not used for training the model), with a sensitivity of 77.46% and a specificity of 80.46%.

Conclusions: The application of artificial intelligence models based on machine learning to classify ear disease amongst Indigenous children is feasible and can achieve an accuracy of more than 80%. The model has not been externally validated. The performance of this model is unlikely to exceed that of an ENT specialist and is therefore more likely to be useful for other health disciplines that have a key role in the delivery of primary care in Indigenous communities. The findings from this study are encouraging of further research and development of artificial intelligence models for the detection of ear disease amongst Aboriginal and Torres Strait Islander children.

Keywords: Telehealth; artificial intelligence (AI); Indigenous health services; ear disease; Aboriginal and Torres Strait Islander

---

^ ORCID: Adam Mothershaw, 0000-0002-0056-8187; Anthony C. Smith, 0000-0002-7756-5136; Liam J. Caffery, 0000-0003-1899-7534.

Introduction

In Australia, otitis media affects Aboriginal and Torres Strait Islander children (hereafter referred to respectfully as Indigenous children) with earlier onset, longer duration, and increased prevalence compared to non-Indigenous children. This disparity is notably true for severe ear conditions (chronic suppurative otitis media, chronic otitis media, perforation) which are almost five times more prevalent amongst Indigenous children (1). If left untreated, infections of the middle ear can cause more serious diseases or hearing loss. One in thirty Indigenous children suffers from hearing loss, most commonly due to chronic otitis media (2). Hearing loss negatively impacts the development and education of children, creating social and economic disadvantage.

For many Indigenous children, the increased likelihood of ear disease is confounded by limited access to specialist healthcare services. In 2015, the full-time equivalent number of medical practitioners per 100,000 population working in major cities was 442, compared to just 279 in outer regional areas and 263 in remote and very remote areas (3). Exacerbating this, specialists are a proportionally smaller segment of the medical practitioner community in outer regional and remote areas. These challenges of distance, remoteness, and practitioner distribution have in part been mitigated by increasing the use of telehealth (4).

Telehealth services which screen for ear disease rely on pre-recorded video-otoscopic images which are transmitted for remote diagnosis (5). There is the potential that artificial intelligence (AI) algorithms can be used to predict ear disease from the information contained in these images. Similar application of AI used to successfully detect ear disease has been previously reported (6,7). However, AI is highly context specific, and there are no studies to date, looking specifically at screening and diagnosis of ear disease amongst Indigenous children. Hence, the aim of this study was to establish proof-of-concept of both the feasibility and effect of using AI to detect ear disease in a telehealth ear screening service for Indigenous children. We present the following article in accordance with the TREND reporting checklist (available at https://dx.doi.org/10.21037/ajo-21-14).

Methods

Setting

Cherbourg is an Aboriginal community located in Queensland, Australia. It is 260 km by road from Queensland's capital city, Brisbane. Around 99% of the town's population of 1,269 are Aboriginal and/or Torres Strait Islander (8). The neighbouring town of Murgon, with a population of 2,378 include approximately 19% Aboriginal and Torres Strait Islander people.

To improve the ear health of children in and around Cherbourg and Murgon, a telehealth-supported mobile Ear, Nose and Throat (ENT) screening service called Health-e-Screen4kids was established in 2009 and has been continually operating since (5,9-12). Ear health screenings are performed by an Indigenous health worker who acquires video-otoscopic images and performs tympanometry. In cases of suspected ear disease, the health worker will refer cases to the local medical service. The referral follows a store-and-forward model of telehealth, and assessments are performed using a secure web interface. When ear disease is confirmed the child is referred to a local general practitioner (GP) or specialist ENT outreach clinic for management.

AI algorithms

Image classifiers are an application of AI that have been used extensively in health to predict or diagnose disease from medical images (13). Large image datasets are a necessary for the training of AI image classifiers using deep learning. Image classification is most frequently implemented using convolutional neural networks (CNN). These have an input layer (the pixels comprising an image), an output layer (disease prediction or diagnosis), and many interconnected hidden layers within. Each hidden layer is composed of a set of mathematical calculations that each take weighted inputs from the previous layer and add a bias before passing an output to the next layer. Training a CNN is the process of adjusting these weights and biases so that the CNN inputs will produce particular network outputs. This process is known as hyperparameter tuning.

Ten years of screening by the Health-e-Screen4Kids service has resulted in a large repository of video-otoscope images of both healthy (no abnormality detected) and abnormal ears, accompanied by patient and assessment metadata including a diagnosis from the ENT specialist's remote review. Labelled images from the Health-e-Screen4Kids service were used to train and validate a binary image classifier model that predicts the likelihood of an input image being either normal or abnormal. Workflows of the screening service meant that images could be labelled with a specific disease condition or alternatively, categorised as abnormal (without specific disease condition).

The architecture for the model was Inception-V3, pre-trained on the ImageNet-1K dataset. The training, validation, and testing was performed with code adapted from Young *et al.* (14). Their study used images of skin lesions to train a binary image classifier model to detect between normal and abnormal skin lesions. The use of Bayesian optimisation with Gaussian processes for hyperparameter tuning in their model was duplicated in our model. The hyperparameter search was run for three trials, each with 30 iterations, and the best-performing (by validation accuracy) model was selected for assessment with the test set.

Dataset preparation

The video-otoscopic images (n=26,033) were retrieved from the clinical database in their original format. The associated assessments from both Indigenous health worker and specialist and additional metadata were also extracted from the database. Patient-identifying data was excluded from the data extraction, ensuring confidentiality.

Pre-processing

Several different models of video-otoscope were used to acquire images, which consequently varied in both file format (JPEG, PNG) and pixel dimensions. Common to all images was a dark empty region outside of the central image circle. Some images had text imprinted in this empty region containing the date, time, or patient name. The empty region and text was removed by cropping every image to the outer boundaries of the image circle. After cropping, all images were resized to a resolution of 299×299 pixels which is the input resolution required by the Inception-V3 model. Finally, the portable network graphic (PNG) images were converted to Joint Photographic Experts Group (JPEG) at 80% quality to standardise the image format.

Labelling

The health worker's screening assessments contained a record of observed abnormalities such as perforation, fluid, retraction, and inflammation. The specialist assessment contained a specific diagnosis, and a grading of the otoscope image quality. This assessment data was used to assign a label to each image as normal or abnormal, or exclude the image from the study. Video-otoscopic images from 923 encounters were excluded due to excessive cerumen

occluding the tympanic membrane. This process yielded 12,742 normal and 2,456 abnormal images. The imbalanced class size was addressed by randomly under sampling the normal images to produce two classes of equal size, thus avoiding a bias towards the majority class. The final dataset contained 8,486 images, evenly split between normal and abnormal. From this dataset 6,818 images were allocated to a training and validation dataset, and the remaining 1,668 images (around 20%) were reserved as the test set. Many of the screening encounters produced multiple images of the same ear canal, usually differing only by a slight change in angle or depth of the video-otoscope. Care was taken to not split these groups between the different sets, as the visual similarity would compromise the test set's purpose of being completely unseen data.

Statistical analysis

Patient characteristic have been analysed and reported using descriptive statistics. The performance of the AI model was analysed and reported using accuracy, sensitivity, and specificity. Accuracy was calculated as the percentage of responses where the output from the image classifier matched the label from the test set. Sensitivity was calculated as the ratio of true positive assessments to all positive assessments (positive assessments includes both true positive and false negatives). Specificity was calculated as ratio of true negative assessments to all negative assessments (negative assessment includes both true negative and false positives).

Ethical statement

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Ethics approval was obtained from Children's Health Queensland Hospital and Health Service Human Research Ethics Committee (No. HREC/06/QRCH/66) and all patients provided informed consent.

Results

The 4,562 patients registered to the program were 52% male and 48% female, with a median age of 6.9±3.64 years at time of first screening. The median age at discharge from the program was 10.9±3.08 years. The Indigenous status of 2,551 patients was recorded, comprising 71% Aboriginal, 1% Torres Strait Islander, 2% Aboriginal and Torres Strait

Table 1 Disease conditions found in the test set and associated accuracy of binary classifier

Disease condition	Number of images with condition in test set	Number correctly classified	Accuracy
Acute otitis media	3	3	100.00%
Chronic suppurative otitis media	14	11	78.57%
Dry ear	6	6	100.00%
Wet ear	13	12	92.31%
Early cholesteatoma	4	3	75.00%
Grommet	39	36	91.31%
Otitis media with effusion	235	189	80.43%
Perforation	34	33	97.06%
Retraction	159	122	76.73%
Abnormal without specific disease condition label	381	278	72.97%
Normal	834	671	80.46%

Islander, and 26% non-Indigenous patients.

The best-performing model achieved an overall accuracy of 80.99% on the validation set and 78.90% on the test set, with a sensitivity of 77.46% and a specificity of 80.46%.

Table 1 shows the distribution of these conditions in the test set, and the accuracy of the binary classifier with those images. This distribution is similar to that of the training and validation sets. As there can be multiple diseases present in an ear, an image may contain multiple labels. Therefore, the total count provided in the table exceeds the number of abnormal test set images.

Discussion

Other studies have explored the use of AI models based on machine learning for automated assessment of video-otoscope images (*Table 2*). These studies report accuracy ranging from 76% to over 99%. Our observed accuracy 80% is generally lower than other studies apart from the 76% reported by Habib *et al.* (15). Differences in accuracy may in part be attributable to the different contexts as few of the other studies used a paediatric dataset and ours was the only one to use a dataset of predominantly Indigenous Australian children. CNNs are complex visual feature extractors, and may therefore be sensitive to the physical differences between adult ears and paediatric ears, which continue to change in size and shape until the age of nine (19).

Accuracy of the model compared to different health disciplines may inform how a model could be deployed in clinical practice. One study reported that an ENT specialist could perform binary classification with an accuracy of between 93% and 100% whereas paediatricians had a slightly lower accuracy of 89% to 100% (20). Again, caution should be applied when using these findings due to contextual differences between the published study and our unique setting. To the best of our knowledge there are no studies reporting accuracy of diagnosis of ear disease in Indigenous children. Nor are there any accuracy studies for other disciplines involved in ear health. Our findings would indicate that the value of an AI model for ENT specialists would be limited. Similarly, other studies have shown that skin disease image classifiers have limited value for experienced dermatologists (21). However, there may be potential for primary care (e.g., GPs or health workers) to triage patients for ENT specialist referral.

Limitations

There are limitations of the dataset that may limit the accuracy of the resulting model. The ground truth label is based on the assessment of a single ENT specialist and potentially some images in the dataset may have been mislabelled. Resultantly, this may potentially confuse the training process and impact the accuracy of the model. This can be mitigated by using more than one person to verify the label before training. The inclusion of only high-quality images only would not reflect clinical reality and the accuracy of our model may be over-reported if used in

Table 2 Artificial intelligence algorithm employing machine learning for the diagnosis of ear disease

Study, year, country	Application	Dataset	Accuracy
Habib *et al.* 2020, Australia (15)	Binary classification of intact or perforated TM	Images sourced from Google images. Training (n=183); test (n=50); intact TM (n=105); perforated TM (n=128); patient demographics not stated; high quality images only included in data set	Overall 76.0%; small perforation size 85.7%; medium perforation size 85.7%; large perforation size 63.6%
Viscaino *et al.* 2020, Chile (16)	Multi-disease classifier; 4 classes (normal, earwax, myringosclerosis, chronic otitis media)	Total (n=720), 80% training an 20% validation; equal number of images in each class	99.03%
Lee *et al.* 2019, Korea (17)	Binary classifier presence or absence of perforation	Training (n=1,338): normal (n=714) and perforation (n=624); validation (n=1,818): normal (n=1,436 and perforation (n=382)	91.0%
Başaran *et al.* 2019, Turkey (7)	Binary classification dataset containing normal and abnormal images (AOM, earwax, Myringosclerosis, tympanostomy tubes, CSOM, otitis externa)	Normal (n=154); abnormal (n=128); after augmentation normal (n=925); abnormal (n=768); patients (n=282); age (range, 2–71 years, mean =8 years); high quality images only included in dataset	90.48%
Cha *et al.* 2019, Korea (6)	Ensemble multi-disease classifier; 6 classes (normal, attic retraction, TM perforation, otitis externa ± myringitis, tumour)	Total (n=10,544 images) 80% training and 20% validation; normal (n=4,342); abnormal (n=6,202); high quality images only included in data set	Ensemble 93.67%; accuracy of individual models (range 85.55–91.55%)
Myburgh *et al.* 2018, South Africa (18)	Multi-disease classifier; 5 classes (normal, wax or FB; AOM, OME, CSOM)	Total (n=389) 80% training and 20% validation; normal (n=123); abnormal (n=266); high quality images only	86.84%

TM, tympanic membrane; AOM, acute otitis medica; CSOM, chronic supportive otitis media; FB, foreign body; OME, otitis media with effusion.

a real-world situation. Furthermore, all ear images used in the training and testing of the AI model were from a single service. Therefore, there was no external validation of the model's accuracy. Hence, the findings from this study are limited to proof-of-concept and technical feasibility. Proof-of-concept studies cannot be used to validate a model's real-world clinical performance (22). The accuracy of the model is likely to improve with a larger training set that include images from beyond the subject service. The amount of data needed to train AI models is dependent on the complexity of the diagnosis. Tasks that are easy to solve for a human reader requires less training data than the detection of subtle or uncommon pathologies (23).

As expected from the screening of an asymptomatic population, our dataset predominantly consisted normal ear images. Very imbalanced datasets can often cause the model to have a bias towards the majority class, reducing the accuracy of the minority class (24). For this paper, class balance was achieved by random culling of the normal dataset. While this solves imbalance, it does discard potentially useful data. There are more sophisticated techniques available to address imbalance, and these can apply to both the dataset and the model parameters (such as error calculation) (24). There is potential for future work to explore the value of these techniques with a typically imbalanced screening service dataset.

Ours was a binary classification model and as such only concerned only with the distinction between normal and abnormal ears. The abnormal images comprise several different conditions of varying severity. The overall accuracy does not necessarily represent the accuracy for any one specific condition. This is especially true when the training/validation and test datasets have been prepared without consideration of the specific conditions. The consequence of random distribution of conditions among datasets, combined with their varying appearance and rates of occurrence, causes inconsistent accuracy of specific condition detection. This is of greatest concern

with a condition such as early cholesteatoma, which occurs very rarely but has serious consequences (hearing loss, mastoiditis, meningitis) if left undetected and untreated (25).

While multiclass and multi-label classification would provide greater insight into the training and detection of individual conditions, it does not address the challenge of training classes with very few samples.

Conclusions

Our study demonstrated that the application of AI models based on machine learning to classify ear disease in Indigenous children is feasible and can achieve an accuracy of more than 80%. The model has not been externally validated. Whilst it is unlikely that an AI model will be superior to the diagnostic skills of an experienced ENT specialist, the use of AI could be useful for other health disciplines that have a key role in the delivery of primary care in Indigenous communities. The findings from this study are encouraging of further research and development of AI models for the detection of ear disease amongst Aboriginal and Torres Strait Islander children, and subsequent prospective testing of these models in a range of real-world settings.

Acknowledgments

Footnote

Reporting Checklist: The authors have completed the TREND reporting checklist. Available at https://dx.doi.org/10.21037/ajo-21-14

Data Sharing Statement: Available at https://dx.doi.org/10.21037/ajo-21-14

Peer Review File: Available at https://dx.doi.org/10.21037/ajo-21-14

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/ajo-21-14). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Ethics approval was obtained from Children's Health Queensland Hospital and Health Service Human Research Ethics Committee (No. HREC/06/QRCH/66) and all patients provided informed consent.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Gunasekera H, Knox S, Morris P, et al. The spectrum and management of otitis media in Australian indigenous and nonindigenous children: a national study. Pediatr Infect Dis J 2007;26:689-92.
2. Australian Bureau of Statistics Australian Aboriginal and Torres Strait Islander Health Survey: First Results, Australia, 2012-13. 2013. Available online: https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/0BBD25C6FF8BDB06CA257C2F001458BF?opendocument (Accessed: 9 June 2021).
3. Australian Institute of Health Welfare Medical practitioners workforce 2015 2016 Available online: https://www.aihw.gov.au/reports/workforce/medical-practitioners-workforce-2015 (accessed: 9 June 2021).
4. Bradford NK, Caffery LJ, Smith AC. Telehealth services in rural and remote Australia: a systematic review of models of care and factors influencing success and sustainability. Rural Remote Health 2016;16:3808.
5. Smith AC, Brown C, Bradford N, et al. Monitoring ear health through a telemedicine-supported health screening service in Queensland. J Telemed Telecare 2015;21:427-30.
6. Cha D, Pae C, Seong SB, et al. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. EBioMedicine 2019;45:606-14.
7. Başarana E, Cömertb Z, Çelik Y. Convolutional neural network approach for automatic tympanic membrane detection and classification. Biomed Signal Process

Control 2020;56:101734.

8. Australian Bureau of Statistics 2016 Census QuickStats - Cherbourg 2016 Available online: https://quickstats.censusdata.abs.gov.au/census_services/getproduct/census/2016/quickstat/UCL315016?opendocument (accessed: 9 June 2021).

9. Elliott G, Smith AC, Bensink ME, et al. The feasibility of a community-based mobile telehealth screening service for Aboriginal and Torres Strait Islander children in Australia. Telemed J E Health 2010;16:950-6.

10. Smith AC, Armfield NR, Wu WI, et al. A mobile telemedicine-enabled ear screening service for Indigenous children in Queensland: activity and outcomes in the first three years. J Telemed Telecare 2012;18:485-9.

11. Nguyen KH, Smith AC, Armfield NR, et al. Cost-Effectiveness Analysis of a Mobile Ear Screening and Surveillance Service versus an Outreach Screening, Surveillance and Surgical Service for Indigenous Children in Australia. PLoS One 2015;10:e0138369.

12. Nguyen KH, Smith AC, Armfield NR, et al. Correction: Cost-effectiveness analysis of a mobile ear screening and surveillance service versus an outreach screening, surveillance and surgical service for indigenous children in Australia. PLoS One 2020;15:e0234021.

13. Wang W, Liang D, Chen Q, et al. Medical Image Classification Using Deep Learning. In: Chen YW, Jain LC. editors. Deep Learning in Healthcare: Paradigms and Applications. Cham: Springer International Publishing, 2020:33-51.

14. Young K, Booth G, Simpson B, et al. Deep neural network or dermatologist? In: Suzuki K, Reyes M, Syeda-Mahmood T, et al. editors. Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. Springer, 2019:48-55.

15. Viscaino M, Maass JC, Delano PH, et al. Computer-aided diagnosis of external and middle ear conditions: A machine

learning approach. PLoS One 2020;15:e0229226.

16. Lee JY, Choi SH and Chung JW. Automated Classification of the Tympanic Membrane Using a Convolutional Neural Network. Appl Sci 2019;9:1827.

17. Myburgh HC, Jose S, Swanepoel DW, et al. Towards low cost automated smartphone- and cloud-based otitis media diagnosis. Biomed Signal Process Control 2018;39:34-52.

18. Habib AR, Wong E, Sacks R, et al. Artificial intelligence to detect tympanic membrane perforations. J Laryngol Otol 2020;134:311-5.

19. Wright CG. Development of the human external ear. J Am Acad Audiol 1997;8:379-82.

20. Pichichero ME, Poole MD. Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. Arch Pediatr Adolesc Med 2001;155:1137-42.

21. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. Nat Med 2020;26:1229-34.

22. Kim DW, Jang HY, Kim KW, et al. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. Korean J Radiol 2019;20:405-10.

23. Weikert T, Francone M, Abbara S, et al. Machine learning in cardiovascular radiology: ESCR position statement on design requirements, quality assessment, current applications, opportunities, and challenges. Eur Radiol 2021;31:3909-22.

24. Wang SJ, Liu W, Wu J, et al. Training Deep Neural Networks on Imbalanced Data Sets. 2016 International Joint Conference on Neural Networks (IJCNN), 2016:4368-74.

25. Holt JJ. Cholesteatoma and otosclerosis: two slowly progressive causes of hearing loss treatable through corrective surgery. Clin Med Res 2003;1:151-4.