Evaluating the generalizability of deep learning image classification algorithms to detect middle ear disease using otoscopy

Al-Rahim Habib (arhabib9@gmail.com)
University of Sydney

Yixi Xu
Microsoft (United States)

Kris Bock
Azure FastTrack Engineering

Shrestha Mohanty
Microsoft (United States)

Tina Sederholm
Microsoft (United States)

William B. Weeks
Microsoft (United States)

Rahul Dodhia
Microsoft (United States)

Juan Lavista Ferres
Microsoft (United States)

Chris Perry
University of Queensland Medical School

Raymond Sacks
University of Sydney

Narinder Singh
University of Sydney

Article

Keywords:

Posted Date: September 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2014320/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Purpose

To evaluate the generalizability of artificial intelligence (AI)-otoscopy algorithms to identify middle ear disease using otoscopic images.

Methods

1842 otoscopic images were collected from 3 independent sources: a) Van, Turkey, b) Santiago, Chile, and c) Ohio, USA. Diagnostic categories consisted of (i) normal or (ii) abnormal. Deep and transfer learning-based methods were used to develop models to evaluate internal and external performance, using area under the curve (AUC) estimates. A pooled assessment was performed by combining all cohorts together with 5-fold cross validation.

Results

AI-otoscopy algorithms achieved high internal performance (mean AUC: 0.95, 95%CI: 0.80–1.00). However, performance was reduced when tested on external otoscopic images not used for training (mean AUC: 0.76, 95%CI: 0.61–0.91). Overall, external performance was significantly lower than internal performance (mean difference in AUC: -0.19, mean standard error: 0.02, p≤0.04). Combining cohorts achieved a substantial pooled performance (AUC: 0.96, standard error: 0.01).

Conclusion

Internally applied AI-otoscopy algorithms performed well in identifying middle ear disease from otoscopy images. However, external performance was reduced when applied to new test cohorts. Further efforts are required to explore data augmentation and pre-processing techniques that might improve external performance and develop a robust, generalizable algorithm for real-world clinical applications.

Introduction

It estimated that over 1.5 billion people in the world live with hearing loss, representing nearly 20% of the global population. The prevalence of hearing loss is expected to rise to over 2.5 billion by 2050. In the US, it is estimated that the overall cost of deafness and hearing loss amounts to US$980 billion annually. Direct and indirect costs related to hearing loss are comprised of health sector costs, educational support, loss of productivity, and societal costs. In the global pediatric population, 34 million children experience hearing loss, of which, 60% can be attributed to preventable causes. The WHO estimates that over one-third of preventable childhood hearing loss is attributed to infections including mumps, rubella, meningitis, measles and middle ear infections.

Otoscopy is an important component of ear health and hearing assessments to identify ear infections. Otoscopy is performed routinely by multiple healthcare workers including medical students, nurses, general practitioners, emergency medicine physicians, paediatricians, audiologists, and otolaryngologists. However, otoscopic interpretation accuracy varies by user experience. Pichichero et al. (2005) demonstrated differences between general practitioners, paediatricians, and otolaryngologists in accurately recognising tympanic membrane (TM) abnormalities. Significant performance differences were found when non-otolaryngologists were asked to differentiate between acute otitis media (AOM), otitis media with effusion (OME), or retracted TMs. These findings are consistent with recent studies that demonstrate that diagnostic accuracy is reduced when non-experts must differentiate between multiple ear disease sub-types. Otolaryngologists significantly outperform non-otolaryngologists in identifying ear disease sub-types and important pathological characteristics (e.g. tympanic membrane perforations, attic retraction, or myringitis).
Autonomous image classification algorithms for otoscopy have been developed using machine learning and deep learning techniques. The purpose of developing artificial intelligence (AI)-based algorithms for otoscopy is to bridge the gap between expert and non-expert performance to improve early ear disease detection. In a recent systematic review and meta-analysis, binary classification algorithms achieved 90.7% accuracy in differentiating between normal versus abnormal otoscopic images. Multi-classification algorithms (i.e., 3 or more diagnostic classes) achieved 97.6% accuracy. However, substantial heterogeneity was found between results proposed using binary and multi-classification algorithms. Although existing literature demonstrates the potential for AI to augment otoscopy and contribute to autonomous classification in the future, substantial heterogeneity between existing algorithms suggests performance may not be generalisable to new test environments. The inability to translate performance in virtual training environments to real-world clinical applications could limit the integration, reliability, and sustainability of AI-based algorithms for otoscopy. Variability in data source, capture devices, ear disease sub-type, and deep learning methods may impact classification performance in clinical encounters different to those used for training. To date, an exploration of the generalisability of AI-based algorithms for otoscopy has not been evaluated.

The purpose of this study was to construct an image classifier capable of differentiating between normal and abnormal otoscopic images using 3 independently collected otoscopy databases, evaluate the generalisability of the algorithm on images not used for training, and explore the optimal convolutional neural network (CNN) and deep learning approach to optimise performance.

Methods

Ethics

This research study was conducted in accordance with the Helsinki Declaration, the Standard for Reporting Diagnostic Accuracy Studies (STARD) and the Consolidated Standards of Reporting Trials for interventions involving artificial intelligence (CONSORT-AI) reporting guidelines. This study utilised 3 open-access otoscopic image databases that were collected after obtaining Institutional Review Board (IRB) approval from each of the respective human research ethics committees and published online for public access, as indicated by each of the authorship groups (Bitlis Eren University Ethics Committee, University of Chile Scientific and Research Ethics Committee, Ohio State University Institutional Review Board).

Data Sources

Figure 1 (A-1 to C-2) summarizes a collection of otoscopic images used in this study. In Turkey (Cohort A), 848 otoscopic images were collected at the Özel Van Akdamar Hospital using digital video otoscopes (unspecified brand and model) with ground-truth classified by 2 otolaryngologists and 1 paediatrician as: normal (Fig. 1A-1), acute otitis media (AOM) (Fig. 1A-2), and chronic otitis media (COM) (Fig. 1A-3, Fig. 1A-4) consisting of chronic suppurative otitis media. In Chile (Cohort B), 540 otoscopic images were collected at the University of Chile Clinical Hospital using the Firefly DE500 digital video otoscope (Firefly Global, Belmont, MA, US. 2022) with ground-truth classified by 1 otolaryngologist as: normal (Fig. 1B-1) or COM (Fig. 1B-2, Fig. 1B-3). In the United States (Cohort C), 454 otoscopic images were collected at Nationwide Children's Hospital and Ohio State University using the JEDMED HORUS+ (JEDMED®, St Louis, MO, US, 2022) with ground-truth classified by 1 otolaryngologist as: normal (Fig. 1C-1) or otitis media with effusion (OME) (Fig. 1C-2). Otoscopic images where the TM could not be visualised were excluded (e.g., obstruction with wax or foreign body).

Legend

Cohort A (ÖZel Van Akdamar Hospital, Van, Turkey): A-1 – normal, A-2 – acute otitis media, A-3 and A-4 – chronic otitis media

Cohort B (University of Chile's Hospital, Santiago, Chile): B-1 – normal, B-2 and B-3 – chronic otitis media

Cohort C (Nationwide Children's Hospital, Columbus, Ohio, US): C-1 – normal, C-2 – otitis media with effusion

Algorithm development
Deep learning-based binary class algorithms were developed to classify normal or abnormal otoscopic images using Cohort A, B and C independently. Abnormal sub-classes included AOM, OME, and COM. Pre-existing ground-truth labels were available for each otoscopic image from the online data sources.\textsuperscript{17–19}

Image classification algorithms were generated using deep and transfer learning to harness prior knowledge gained by CNNs from previous classification tasks.\textsuperscript{20,21} The final classification layer of the CNNs were fine-tuned to yield a custom model developed specifically for otoscopic images. Pre-trained CNNs were selected based on results from previously published experiments demonstrating superior performance in classifying otoscopic images.\textsuperscript{17,22–25} Pre-trained CNNs included: ResNet-50\textsuperscript{26}, VGGNet-16\textsuperscript{27}, DenseNet-161\textsuperscript{28}, and Vision Transformer.\textsuperscript{29} Images were resized to 224 by 224 pixels and trained over 200 epochs with a learning rate of 0.0001 using the Microsoft Azure™ Machine Learning Studio (Redmond, Washington, USA). Cross-validation utilised stratified sampling across 5 folds of which 3 were used for training, 1 for validation, and 1 for testing.

A summary of models is provided in Table 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Data source</th>
<th>Classes</th>
<th>Abnormal subtype(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cohort A (Özel Van Akdamar Hospital, Turkey)</td>
<td>Normal, Abnormal</td>
<td>AOM, COM</td>
</tr>
<tr>
<td>2</td>
<td>Cohort B (University of Chile Clinical Hospital, Chile)</td>
<td>Normal, Abnormal</td>
<td>COM</td>
</tr>
<tr>
<td>3</td>
<td>Cohort C (Nationwide Children's Hospital and Ohio State University, Ohio, US)</td>
<td>Normal, Abnormal</td>
<td>OME</td>
</tr>
<tr>
<td>4</td>
<td>Combined database (Cohort A, B and C)</td>
<td>Normal, Abnormal</td>
<td>AOM, COM, OME</td>
</tr>
</tbody>
</table>


Statistical analysis

To compare the algorithm to the reference standard for Cohorts A, B, and C, contingency tables were generated to summarise Area Under the Curve (AUC), accuracy, sensitivity, and specificity. Confusion matrices were constructed to summarise correct and incorrect predictions across each model and test cohort. To evaluate internal performance, models were trained and tested using otoscopic images from the same cohort (e.g., split Cohort A otoscopic images into training and test sets). To evaluate external validation performance and generalizability, models were validated on external cohorts (e.g., trained using Cohort A and validated on Cohort B and Cohort C).\textsuperscript{30} Mean difference in AUC were calculated to compare internal and external performance. The student’s t-test was used to evaluate the null hypothesis that internal and external performance were equal. Pooled assessment was performed by combining Cohort A, B, and C using split datasets for training, validation, and 5-fold cross validation was applied.

Statistical analysis was performed using Stata 17 (StataCorp LLC, College Station, Texas, USA, 2021).

Results

In total, 1842 otoscopic images were collected from 3 online resources, comprised of 1117 normal and 725 abnormal images. Table 2 summarises the distribution of normal and abnormal otoscopic images by Cohort A, B and C.
### Table 2
Summary of data sources

<table>
<thead>
<tr>
<th>Otoscope</th>
<th>Image resolution (pixels)</th>
<th>Ground-truth</th>
<th>Total Images</th>
<th>Normal Images</th>
<th>Abnormal Images</th>
<th>Ear disease diagnoses</th>
<th>Published accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort A (Özel Van Akdamar Hospital, Van, Turkey)</td>
<td>Unspecified brand / model</td>
<td>500 x 500 (DPI: 72)</td>
<td>2 otolaryngologists, 1 paediatrician</td>
<td>848</td>
<td>578</td>
<td>270</td>
<td>Normal, AOM, COM</td>
</tr>
<tr>
<td>Cohort B (University of Chile's Hospital, Santiago, Chile)</td>
<td>Firefly DE500 digital video otoscope</td>
<td>420 x 380 (DPI: 72)</td>
<td>1 otolaryngologist</td>
<td>540</td>
<td>360</td>
<td>180</td>
<td>Normal, COM</td>
</tr>
<tr>
<td>Cohort C (Nationwide Children's Hospital, Columbus, Ohio, US)</td>
<td>JEDMED HORUS+ digital video otoscope</td>
<td>1397 x 982 to 441 x 355 (DPI: 72)</td>
<td>1 otolaryngologist</td>
<td>454</td>
<td>179</td>
<td>275</td>
<td>Normal, OME</td>
</tr>
<tr>
<td>Combined (Cohort A, B, C)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1842</td>
<td>1117</td>
<td>725</td>
<td>Normal, AOM, OME, COM</td>
</tr>
</tbody>
</table>


### Internal performance

Table 3 summarises internal performance for each cohort. Models trained using Cohort A otoscopic images achieved AUC between 0.82 to 0.86 (accuracy: 80 to 84%; sensitivity: 57 to 70%; specificity 90 to 91%). Models trained using Cohort B otoscopic images achieved an AUC of 1.00 (100% accuracy, 100% sensitivity, and 100% specificity). Models trained using Cohort C otoscopic images achieved AUC between 0.98 to 0.99 (accuracy: 91 to 95%; sensitivity: 91 to 96%; specificity: 89 to 96%).
### Table 3
Summary of internal and external performance between Cohorts to differentiate normal versus abnormal otoscopic images trained using DenseNet-161

<table>
<thead>
<tr>
<th>Experiments</th>
<th>DenseNet-161</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>Validate</td>
<td>AUC</td>
<td>Acc (%)</td>
<td>Sen (%)</td>
</tr>
<tr>
<td>Cohort A</td>
<td>Cohort A</td>
<td>0.86</td>
<td>84</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>(Internal)</td>
<td>(95%CI: 0.8,0.93)</td>
<td>(95%CI: 78,89)</td>
<td>(95%CI: 56,82)</td>
</tr>
<tr>
<td>Cohort B</td>
<td>Cohort B</td>
<td>0.80</td>
<td>62</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>(External)</td>
<td>(95%CI: 0.76,0.84)</td>
<td>(95%CI: 58,66)</td>
<td>(95%CI: 87,95)</td>
</tr>
<tr>
<td>Cohort C</td>
<td>Cohort C</td>
<td>0.72</td>
<td>66</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>(External)</td>
<td>(95%CI: 0.68,0.77)</td>
<td>(95%CI: 61,70)</td>
<td>(95%CI: 88,95)</td>
</tr>
<tr>
<td>Cohort B</td>
<td>Cohort B</td>
<td>1.00</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>(Internal)</td>
<td>(95%CI: 1,1)</td>
<td>(95%CI: 97,100)</td>
<td>(95%CI: 90,100)</td>
</tr>
<tr>
<td>Cohort A</td>
<td>Cohort A</td>
<td>0.61</td>
<td>65</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>(External)</td>
<td>(95%CI: 0.57,0.65)</td>
<td>(95%CI: 62,68)</td>
<td>(95%CI: 36,48)</td>
</tr>
<tr>
<td>Cohort C</td>
<td>Cohort C</td>
<td>0.90</td>
<td>80</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>(External)</td>
<td>(95%CI: 0.88,0.93)</td>
<td>(95%CI: 76,84)</td>
<td>(95%CI: 67,78)</td>
</tr>
<tr>
<td>Cohort C</td>
<td>Cohort C</td>
<td>0.98</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>(Internal)</td>
<td>(95%CI: 0.97,1)</td>
<td>(95%CI: 83,96)</td>
<td>(95%CI: 80,97)</td>
</tr>
<tr>
<td>Cohort A</td>
<td>Cohort A</td>
<td>0.62</td>
<td>48</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>(External)</td>
<td>(95%CI: 0.58,0.66)</td>
<td>(95%CI: 45,52)</td>
<td>(95%CI: 76,86)</td>
</tr>
<tr>
<td>Cohort B</td>
<td>Cohort B</td>
<td>0.91</td>
<td>35</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>(External)</td>
<td>(95%CI: 0.89,0.93)</td>
<td>(95%CI: 31,40)</td>
<td>(95%CI: 98,100)</td>
</tr>
<tr>
<td>Mean performance all cohorts (Internal)</td>
<td>0.95</td>
<td>92</td>
<td>87</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>(95%CI: 0.86, 1.00)</td>
<td>(95%CI: 83, 100)</td>
<td>(95%CI: 70, 100)</td>
<td>(95%CI: 89, 100)</td>
</tr>
<tr>
<td>Mean performance all cohorts (External)</td>
<td>0.76</td>
<td>59</td>
<td>80</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>(95%CI: 0.61, 0.91)</td>
<td>(95%CI: 42, 77)</td>
<td>(95%CI: 56, 100)</td>
<td>(95%CI: 9, 83)</td>
</tr>
</tbody>
</table>

Legend: Acc – accuracy, AUC – area under the Receiver Operating Characteristics curve, Sen – sensitivity (abnormal), Spec – specificity (normal), 95%CI – 95% confidence interval

Table 3 summarises external validation performance for each cohort. Models trained using Cohort A otoscopic images and tested on Cohort B achieved AUC between 0.80 to 0.87 (accuracy: 62 to 77%; sensitivity: 76 to 96%; specificity: 47 to 78%), while those tested on Cohort C achieved AUC between 0.61 to 0.82 (accuracy: 65 to 76%; sensitivity: 80 to 92%; specificity: 25 to 69%).

Models trained using Cohort B otoscopic images and tested on Cohort A achieved AUC between 0.60 to 0.67 (accuracy: 59 to 65%; sensitivity: 4 to 54%; specificity: 61 to 99%), while those tested on Cohort C achieved AUC between 0.87 to 0.91 (accuracy: 72 to 80%; sensitivity: 62 to 73%; specificity: 88 to 95%) (Table 3).

Models trained using Cohort C otoscopic images and tested on Cohort A achieved AUC: 0.54 to 0.68 (accuracy: 44 to 70%; sensitivity: 14 to 82%; specificity: 26 to 96%), while those tested on Cohort B achieved AUC: 0.85 to 0.94 (accuracy: 34 to 45%; sensitivity: 100%; specificity: 1 to 17%) (Table 3).
On average, external AUC was significantly lower than internal AUC (DenseNet-161 mean difference –AUC: -0.18 ± 0.02, p < 0.01). As shown in Table 4, Cohort A had the smallest difference between internal and external AUC (mean difference: -0.06 to -0.14; SE: 0.02, p < 0.01 to 0.04) compared to Cohort B (mean difference: -0.10 to -0.39; standard error: 0.01 to 0.02; p < 0.01) or Cohort C (mean difference: -0.07 to -0.36; standard error: 0.01 to 0.02; p < 0.01).

Table 4
Summary of mean difference in Area Under the Curve (AUC) estimates between internal and external performance to differentiate normal versus abnormal otoscopic images trained using DenseNet-161

<table>
<thead>
<tr>
<th>DenseNet-161</th>
<th>Train</th>
<th>External</th>
<th>Mean Difference in AUC (Internal – External)</th>
<th>Standard Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort A</td>
<td>Cohort B</td>
<td>-0.06</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Cohort A</td>
<td>Cohort C</td>
<td>-0.14</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>Cohort B</td>
<td>Cohort A</td>
<td>-0.39</td>
<td>-0.40</td>
<td>-0.40</td>
<td></td>
</tr>
<tr>
<td>Cohort B</td>
<td>Cohort C</td>
<td>-0.10</td>
<td>-0.09</td>
<td>-0.09</td>
<td></td>
</tr>
<tr>
<td>Cohort C</td>
<td>Cohort A</td>
<td>-0.36</td>
<td>-0.31</td>
<td>-0.31</td>
<td></td>
</tr>
<tr>
<td>Cohort C</td>
<td>Cohort B</td>
<td>-0.07</td>
<td>-0.14</td>
<td>-0.14</td>
<td></td>
</tr>
</tbody>
</table>

Pooled performance

Models developed using a combination dataset of all cohorts achieved 90 to 91% accuracy (AUC: 0.96; sensitivity: 84 to 87%; specificity: 93 to 95%) (Table 5). DenseNet-161 had the highest AUC and smallest standard deviation for 5-fold cross validation.

Table 5
Summary of pooled performance to differentiate normal versus abnormal otoscopic images using combined dataset (Cohort A, B, and C) by pre-trained CNN architecture

<table>
<thead>
<tr>
<th>Model architecture</th>
<th>Acc (SE)</th>
<th>AUC (SE)</th>
<th>Sen (SE)</th>
<th>Spec (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>0.91 (0.02)</td>
<td>0.96 (0.01)</td>
<td>0.87 (0.04)</td>
<td>0.93 (0.01)</td>
</tr>
<tr>
<td>DenseNet-161</td>
<td>0.91 (0.02)</td>
<td>0.96 (0.01)</td>
<td>0.86 (0.04)</td>
<td>0.94 (0.01)</td>
</tr>
<tr>
<td>VGG16</td>
<td>0.90 (0.02)</td>
<td>0.96 (0.01)</td>
<td>0.84 (0.02)</td>
<td>0.94 (0.02)</td>
</tr>
<tr>
<td>Vision Transformer</td>
<td>0.92 (0.01)</td>
<td>0.96 (0.01)</td>
<td>0.86 (0.01)</td>
<td>0.95 (0.01)</td>
</tr>
</tbody>
</table>


Discussion

Otoscopy is a common clinical task performed by multiple healthcare workers and practitioners with varying levels of expertise. Identifying normal anatomical landmarks and abnormal pathological processes is important in evaluating ear health. The diagnosis of ear disease relies on a combination of presenting symptoms, clinical history, tympanometry, audiometry, and otoscopic findings. AI-based algorithms have been explored in healthcare to augment existing clinical practices in order to enhance judgement and decision-making. Al-based algorithms for otoscopy have achieved substantial accuracy to differentiate between normal or abnormal otoscopic images utilising pre-trained CNN architectures. Despite achieving high performance in training and internal validation groups, substantial heterogeneity has been found between model results. The differences may, in part, be attributed to the variability in image capture methods or devices, definition of diagnoses, ground truth
determination, sample populations, and methodology used for algorithm development. Translating performance achievements from training environments to new test cohorts is an important aspect for an AI-based model to be generalizable for routine clinical practice. Without achieving generalizable performance and real-world utility, it is unclear if applying AI to otoscopy practices will contribute to improving clinical practice. The purpose of this study was to evaluate the generalisability of a binary classification AI-otoscopy algorithm to classify normal or abnormal otoscopic images, using 3 independently collected databases and trained using standardised deep learning methods.

This study demonstrates substantial internal performance within each cohort to differentiate between normal or abnormal otoscopic images. Between CNN architectures, models developed using DenseNet-161 achieved the highest internal performance for the pooled assessment. However, when the models were applied externally (i.e., to new validation cohorts not used for training), overall model performance was reduced. The extent of performance variability was dependent on the outcome measure assessed (e.g., accuracy, AUC, sensitivity, and specificity). Performance was reduced in most external comparisons, although the mean difference varied by outcome measure. Cohort A (Özel Van Akdamar Hospital, Van, Turkey) was found to have the smallest external performance drop in accuracy, AUC, and sensitivity. On the other hand, Cohort B (University of Chile’s Hospital, Santiago, Chile) was found to have the smallest performance drop in specificity, but the greatest performance drop in AUC and sensitivity. These findings demonstrate that measuring external model performance and evaluating generalizability to new cohorts need to be interpreted relative to the outcome measure of interest. The outcome variable 'accuracy' evaluated final predictions as discrete categories, where predictions could be considered either correct or incorrect. However, the outcome variable 'AUC' measured the degree of separability and, as previously demonstrated, is a better measure of algorithm performance than accuracy.

External validation performance differences between cohorts also may be attributed to intrinsic differences in the number of otoscopic images available for training, ear disease categories, and image capture devices. Cohort A had the lowest internal performance despite having both the greatest number of otoscopic images available for training and number of ear disease categories within overarching 'abnormal' classification labels. Although these factors may have contributed to less accurate internal performance due to greater heterogeneity in model learning across each category, these deficiencies were not translated to the model's ability to generalize compared to Cohort B and C.

Interestingly, models naïve to OME during training were able to achieve an AUC between 0.72 to 0.90 and to differentiate OME from normal TMs in validation cohorts. This achievement may reflect the CNNs architectural design and the feature extraction layers that were used to extract pertinent features common to normal TMs, ones that are translatable to new cohorts. However, the addition of AOM in Cohort A may have negatively impacted the CNNs ability to extract pertinent features as, often, AOM, OME, and normal TMs are phenotypically comparable and challenging to differentiate, even among experts.

This study represents the first attempt to evaluate the generalisability of AI-otoscopy algorithms trained and tested on 3 unique and independently collected databases. Although the image capture devices and data collection protocols varied by each cohort, the deep learning methodology was standardised and uniform between groups. This approach revealed the potential for pre-trained CNNs to extract important features from otoscopic images used during training that could be applied universally to new images. Comparable to a healthcare worker's evolving knowledge base that improves with otoscopy experience, CNNs demonstrate the potential to identify useful patterns in otoscopic images despite differences in otoscopes, image quality, and underlying ear disease diagnoses. The patterns and features extracted by pre-trained CNNs enable broad differentiation between normal versus abnormal TMs.

This study has several limitations. First, this study utilises secondary data published online for open-access use. As a result, the data collection protocols, image capture devices, ground-truth definitions, and labelling formats were not standardised across cohorts. Although the models developed for this study evaluated normal versus abnormal classifications, it is plausible that differences in the ear disease definitions and potential misclassifications could influence the models' performance and overestimate its accuracy. Second, it is unclear if the ear disease diagnoses were based on otoscopic images alone or in concert with tympanometry, audiometry, pneumatic otoscopy, and clinical history. Inter-cohort variability and interpretation bias may exist in the ground-truth labels, potentially contributing to external measurement errors and misrepresenting the algorithm's generalisability. Third, ground-truth labels were not established by expert consensus. Inter-rater variability and rater bias could impact the overall reliability of ground-truth labels and adversely impact external validation.
The potential for AI-based algorithms to augment otoscopy is promising. Autonomous AI-otoscopy tools have the potential to improve triage and clinical decision-making. In settings where the prevalence of ear disease is high and access to healthcare services is limited (e.g., rural and remote areas; Indigenous communities in North America, Australia, and New Zealand; under-resourced or marginalised populations in sub-Saharan Africa, India, south-east Asia)\textsuperscript{42–44}, AI-otoscopy algorithms may enhance telehealth programs and empower local healthcare workers to perform otoscopy with more certainty. However, for this to be achieved, AI-otoscopy models must be accurate, reliable, generalizable, accessible in online/offline settings, sustainable, and ethnically constructed and deployed. Early studies have demonstrated that CNNs can achieve substantial performance to interpret otoscopy, although, as shown in this study, external validation performance and generalisability to new settings is limited. Further efforts are required to expand otoscopic image databases to include images from multiple countries, rural versus urban settings, various image capture devices, and include users from various experience levels to account for variability in image capture quality. Establishing standardised diagnostic definitions or consensus from expert panels is required to ensure the ground-truth labels are valid and subsequent models are trained with minimal classification bias. The consideration of data augmentation and pre-processing techniques may be explored to enhance external performance and overcome performance limitations. Otoscopy practices will inevitably vary between clinical settings and a robust, comprehensive, and generalizable AI-otoscopy algorithm is necessary for real-world applications.

**Conclusion**

Otoscopy is an important component of ear health and hearing assessments. This study explores internal and external validation performance to evaluate generalisability of AI-otoscopy algorithms. AI-otoscopy algorithms achieve substantial internal performance to identify middle ear disease from otoscopy images. Internal and pooled performance results are comparable to previously reported findings. However, external performance was reduced when applied to new test cohorts. Further efforts are required to explore data augmentation and pre-processing techniques to improve external performance and develop a robust, generalizable algorithm for real-world clinical applications.

**Declarations**

**Acknowledgements**

We acknowledge Microsoft’s AI for Humanitarian Action Grant, the Garnett Passe and Rodney Williams Memorial Foundation, the AI for Good Research Lab, and the Avant Foundation for providing research funding to support this project.

**Author contributions**

A-RH, YX, KB, SM, TS, RD, JLF, CP, RS, NS conceived the study idea. A-RH, YX, KB, SM were involved in data collection. A-RH, YX extracted data and performed the data analysis. A-RH wrote the first draft of the manuscript. A-RH, YX, KB, SM, TS, WBW, RD, JLF, CP, RS, interpreted the data analysis and critically revised the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. All authors read and approved the final manuscript.

**Data availability statement**

Data that support the findings of this study are available from the corresponding author on reasonable request.

**Competing Interests Statement**

A-RH discloses research funding from the Garnett Passe and Rodney Williams Memorial Foundation, Microsoft’s AI for Humanitarian Action grant, and the Avant Mutual Foundation Doctor-in-Training research grant. CP discloses conflict of interest from Australian Medical Association (Queensland) and provides consulting services for the Deadly Ears Program. RS is a consultant for Medtronic. NS discloses research funding from Microsoft’s AI for Humanitarian Action grant, consultant for ResMed, Optinose, Nasus, GSK, and ENT Technologies, and has received conference funding from Medtronic, Karl Storz, and NeilMed. YX, KB, SM, TS, RD, and JLF are affiliated with the AI for Good Research Lab.
References


Figures

Figure 1

(A-1 to C-2). Summary of otoscopic image cohorts used to develop a deep learning algorithm to classify normal versus abnormal otoscopic images, with sample images.