# Digital otoscopy videos versus composite images: A reader study to compare the accuracy of ENT physicians

<table>
<thead>
<tr>
<th>Journal:</th>
<th><em>The Laryngoscope</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>Draft</td>
</tr>
<tr>
<td>Wiley - Manuscript type:</td>
<td>Original Reports</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Binol, Hamidullah; Wake Forest School of Medicine, Niazi, Muhammad Khalid Khan; Wake Forest School of Medicine Essig, Garth; the Ohio State University, Otolaryngology-Head and Neck Surgery Shah, Jay; Case Western Reserve Univ, ENT Mattingly, Jameson; The Ohio State University, Department of Otolaryngology Harris, Michael; Medical College of Wisconsin, Otolaryngology and Communication Sciences Elmaraghy, Charles; Nationwide Children's Hospital, Teknos, Theodoros; Case Comprehensive Cancer Center, Department of Otolaryngology; Taj-Schaal, Nazhat; The Ohio State University, Department of Internal Medicine Yu, Lianbo; the Ohio State University, Department of Biomedical Informatics Gurcan, Metin; Wake Forest School of Medicine Moberly, Aaron; Ohio State University, Otolaryngology</td>
</tr>
<tr>
<td>Keywords - Combo:</td>
<td>Otitis media/ chronic otitis media/conductive hearing loss &lt; Pediatric otology &lt; Pediatrics, Middle ear &lt; Otology/Neurotology, otitis media (OM) &lt; Otology/Neurotology</td>
</tr>
</tbody>
</table>
Title: Digital otoscopy videos versus composite images: A reader study to compare the accuracy of ENT physicians

Running Title: Otoscopy videos vs composite images

Name(s) of the author(s): Hamidullah Binol, PhD¹, M. Khalid Khan Niazi, PhD¹, Garth Essig, MD², Jay Shah, MD³, Jameson K. Mattingly, MD², Michael S. Harris, MD⁴, Charles Elmaraghy, MD², Theodoros Teknos, MD⁵, Nazhat Taj-Schaal, MD⁶, Lianbo Yu, PhD⁷, Metin N. Gurcan, PhD¹, Aaron C. Moberly, MD²

Affiliations of author(s):
¹ Center for Biomedical Informatics, Wake Forest School of Medicine, Winston-Salem, NC, USA
² Department of Otolaryngology, Ohio State University, OH, USA
³ Case Western Reserve University School of Medicine, OH, USA
⁴ Otolaryngology and Comm. Sciences - Froedtert Hospital, WI, USA
⁵ University Hospitals Seidman Cancer Center, OH, USA
⁶ Department of Internal Medicine, Ohio State University College of Medicine, OH, USA
⁷ Department of Biomedical Informatics, Ohio State University, OH, USA

Corresponding author: Hamidullah Binol
Postal address: 486 N. Patterson Avenue, Winston-Salem, NC 27101, USA
Mobile number: (786) 442-5600
Email address: hbinol@wakehealth.edu
Fax: (336) 716-0239
ORCID: https://orcid.org/0000-0003-2537-8280

Funding: The project described was supported in part by Award R21 DC016972 (PIs: Gurcan, Moberly) from National Institute on Deafness and Other Communication Disorders. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Deafness and Other Communication Disorders or the National Institutes of Health.

Conflicts of interest: Authors ACM, GE, and CE are shareholders in Otologic Technologies. Authors ACM and MNG are paid consultants and serve on the Board of Directors for Otologic Technologies.
Abstract

Objectives: With the increasing emphasis on developing effective telemedicine approaches in Otolaryngology, this study explored whether a single composite image stitched from a digital otoscopy video provides acceptable diagnostic information to make an accurate diagnosis, as compared with that provided by the full video.

Methods: Five Ear, Nose, and Throat (ENT) physicians reviewed the same set of 78 digital otoscope eardrum videos from four eardrum conditions: normal, effusion, retraction, and tympanosclerosis, along with the composite images generated by a SelectStitch method that selectively uses video frames with computer-assisted selection, as well as a Stitch method that incorporates all the video frames. Participants provided a diagnosis for each item along with a rating of diagnostic confidence. Diagnostic accuracy for each pathology of SelectStitch was compared with accuracy when reviewing the entire video clip and when reviewing the Stitch image.

Results: There were no significant differences in diagnostic accuracy for physicians reviewing SelectStitch images and full video clips, but both provided better diagnostic accuracy than Stitch images. The inter-reader agreement was moderate.

Conclusion: Equal to using full video clips, composite images of eardrums generated by SelectStitch provided sufficient information for ENTs to make the correct diagnoses for most pathologies. These findings suggest that use of a composite eardrum image may be sufficient for telemedicine approaches to ear diagnosis, eliminating the need for storage and transmission of large video files, along with future applications for improved documentation in electronic medical record systems, patient/family counseling, and clinical training.
Level of Evidence: Level 3

Keywords: Computer-assisted Diagnosis, Eardrum, Image stitching, Otoscope, Telemedicine.

Total words: 2993
Introduction

Clinical examination of the eardrum ( tympanic membrane – TM) through handheld otoscopy is the most common diagnostic approach for TM pathologies \(^1\). Interpretation of the often brief glimpse of the TM obtained through the small viewing window requires extensive experience. With a growing need to develop effective telemedicine systems, which recently gained widespread attention during the COVID-19 pandemic \(^2\), novel methods to perform telemedicine otoscopy are needed. Previous studies have shown that telemedicine review of images is sufficiently accurate to use in burn \(^3\-\^5\) and trauma care \(^6\,\^7\).

For otoscopy, one telemedicine approach is to apply digital otoscopy, during which a short video examination of the TM is recorded, which is reviewed by a telemedicine physician. Although there are no studies directly comparing otoscopic diagnoses based on a video clip as compared with a single digital image, our previous work led us to use videos \(^8\). There, diagnostic performance was compared between otoscopic single images and in-office microscopy. We included only images that were of sufficient focus/lighting, representing relatively ideal imaging conditions. Other authors have also noted insufficient image quality in a large percentage of their otoscopic still image databases, and/or the broad variability inherent across still images \(^9\,\^10\).

The use of digital otoscopic video clips could overcome limitations imposed in real clinical settings: collecting a string of frames in a video could capture at least a few useful frames with sufficient focus and lighting, even in the setting of partially obstructing cerumen or a moving child. However, a major downside of video clips is that they require a large amount of storage space. A typical otoscopic video clip is \(1440 \times 1080 \times 24\) bits/pixels per frame, with between 200 and 1000 frames, contrasted with a single frame
for a still image. Transferring large video clips, even after compression, can provide a major barrier to care in settings where internet bandwidth is limited. Moreover, storage of videos within electronic medical records is not currently streamlined, and replaying videos for patients/families for counseling purposes is tedious.

Along similar lines, previous studies have been done to detect ear abnormalities by computer-based methods, requiring either manually extracting a single image from videos or capturing a single image with minimal glare/obstruction. However, manually selecting a frame from a video is time-consuming and subject to high inter- and intra-reader variability. A more sophisticated computerized method that creates a “composite” otoscopic image from a video should lead to a more useful final image, since a typical video comprises at least 200 frames, more than one of which may contribute information to the diagnostician.

We previously reported a computer-aided otoscopic frame selecting and stitching framework called SelectStitch, in which a semantic segmentation-based framework automatically selects meaningful frames containing portions of the TM from videos, reducing irrelevant frames (e.g., those heavily blurred or having excessive cerumen). We then conducted a reader study with three Ear, Nose, and Throat (ENT) physicians who reviewed these composite SelectStitch images and compared them to composite images generated using the entire video (i.e., without frame selection, called Stitch) in terms of diagnostic decisions. Figure 1 provides an overview of SelectStitch and Stitch. We found that SelectStitch improved the diagnostic quality of composite images relative to Stitch. However, that study did not address several remaining questions important for the
translation of this approach to the clinic and particularly to telemedicine settings, which are addressed in the current study:

(a) When reviewing SelectStitch composite images, what is the accuracy of diagnosis? To answer this question, five ENTs reviewed 78 composite images and provided diagnoses, compared with a “true” diagnosis. For adult patients, the “true” diagnosis was based on digital otoscopy, supplemented with clinical microscopy as well as audiology testing (hearing testing and/or tympanometry). For pediatric patients, the “true” diagnosis was based on digital otoscopy, supplemented with microscopy in the operating room during placement of pressure equalization tubes. We also aimed to determine which pathologies were easiest and hardest to diagnose.

(b) Is the accuracy of diagnosis for SelectStitch composite images different from Stitch images and from videos? Because the Stitch technique generates composite images using all frames of a video, including redundant frames and frames of poor quality, we predicted that the diagnostic accuracy for SelectStitch images would be superior to the accuracy for Stitch. More importantly, we predicted that the diagnostic accuracy for SelectStitch would be equivalent to the accuracy for full video clips.

(c) How does the level of confidence of ENTs for each diagnostic tool (SelectStitch, Stitch, and video) relate to diagnostic ability? To answer this question, the five ENTs rated their level of confidence in making diagnoses for each type of pathology in each diagnostic tool condition.

(d) What is the inter-reader variability of ENTs on diagnosing with the diagnostic tools? As with any medical application, we expected that there would be inter-reader variability among ENTs, but that agreement would generally be relatively high.
Materials and Methods

A database of high-resolution digital adult and pediatric videos, captured via a digital otoscope from ENT clinics and operating rooms, as well as in a primary care Medicine/Pediatrics setting, was created after Institutional Review Board (IRB) approval. A high definition (HD) video otoscope (JEDMED Horus+ HD Video Otoscope, St. Louis, MO) was utilized. The video frames were 1440 by 1080 pixels and were recorded in a MPEG 4 file format. In this study, 78 video clips from the database were used, selected if they only had one single diagnostic label associated. These videos consisted of 20 normal ears, 20 with middle ear effusions (serous or mucoid), 20 with TM retractions, and 18 with tympanosclerosis (i.e., myringosclerosis). Videos were excluded if they had low light throughout the video and/or if they did not contain a clear view of at least part of the TM. Where possible, we selected pediatric and adult videos as balanced (e.g., 10 pediatrics and 10 adults), except for tympanosclerosis for which there were 10 adult and eight pediatric videos.

An online diagnostic assessment tool was designed using SurveyMonkey, an online survey software. The video clips were hosted on Vimeo, and the composite images were uploaded to imgbox. An example of a question from our online survey is shown in Figure 2. Each sample (Stitch or SelectStitch composite image or video) was displayed on the screen, and the reader was asked to state the diagnosis (or normality). The order of presentation (video first or composite image first) was randomized to each clinician separated by four weeks (see Figure 3). If a reader viewed the video of a sample in the first survey, he/she read the Stitch and SelectStitch composite images (also in a randomized
order) in the second survey. The cases from adult and pediatric patients were also mixed in each survey.

At the completion of each survey, readers were asked to rate their degree of confidence in making each type of diagnosis, on a scale of 1 to 5, in which 1 indicated no confidence while 5 indicated extreme confidence. Five ENTs (authors ACM, GE, JS, JKM, and MSH; three neurotologists, one comprehensive otolaryngologist, and one pediatric otolaryngologist) were invited by email to complete the online assessment, and all completed the assessment after completing written informed consent.

Statistical Analyses

Two different scoring strategies were applied to the survey answers. Although each sample had only one true diagnostic label, we did not restrict the readers regarding the number of diagnostic answers they could provide for each sample. Answers were scored using two different strategies. In Score-1, we scored the answers according to whether the reviewer provided the correct diagnostic answer as well as how many answers were given (e.g., the true label was effusion but the reviewer provided two diagnoses: effusion and tympanosclerosis). To compute accuracy using Score-1, an answer-weighting strategy was used: proportion = \( \frac{\delta}{N_A} \) where \( N_A \) was the number of answers provided by the reviewer and \( \delta \) was the binary output of answers, where the item received a 1 if any of the answers were correct and 0 otherwise. For example, if a reader selected two answers (\( N_A = 2 \)) and one of them was correct (\( \delta = 1 \)), then the proportion (in percentage) for that particular sample would be 50%. It should be noted that this is not “accuracy” in a traditional sense. In contrast, for Score-2, an answer was accepted as correct if any diagnosis in the response...
matched the true label. Score-2 was computed as the percent of responses that contained a correct diagnosis, which is a relatively lenient approach to accuracy.

To study differences in scoring among diagnostic tools (video clips, Stitch, SelectStitch) and among the five ENT doctors, ordinal logistic regression was applied to both Score-1 and Score-2. Similar analysis was performed for studying the association between confidence level in scoring and scores. Wald tests were performed for comparison between diagnostic tools. Bonferroni method was used for multiple comparisons (e.g., $\alpha = 0.016$ when adjusting for three comparisons). Kendall’s concordance was calculated to assess inter-reader agreement for each diagnostic tool, where a concordance of 0 suggests no inter-reader agreement and a concordance of 1 suggests perfect agreement.

**Results**

**Question (a): What was the accuracy of diagnosis for SelectStitch?**

As shown in Table 1, the overall proportions of diagnostic accuracy for SelectStitch images among ENTs varied between 46 and 62% for Score-1. The easiest and hardest categories to diagnose, respectively, were Tympanosclerosis (mean ± std: 69% ± 9) and Retraction (mean ± std: 39% ± 7).

For Score-2, also shown in Table 1, the average accuracy rates of ENT doctors for SelectStitch suggested that the easiest category to diagnose was again Tympanosclerosis (mean ± std: 84% ± 7) and the hardest to diagnose was Retraction (mean ± std: 56% ± 12). Overall Score-2 accuracies among ENTs varied between 57 and 74%.
Question (b): Did the accuracy of diagnosis differ among SelectStitch, Stitch, and video clips?

Similar tables of diagnostic accuracy are shown in the Appendix for Stitch (Appendix Table A) and video clips (Appendix Table B). For Score-1, overall, there was a significant difference in score among the three diagnostic tools at p value < 0.0001 (F value = 44.42). For paired comparisons, there was no significant difference between the video method and SelectStitch at p value = 0.9736; there was a significant difference between video method and Stitch (Stitch method scored less) at p value < 0.0001; and there was a significant difference between SelectStitch and Stitch (Stitch scored less) at p value < 0.0001 (Table 2). For Score-2, overall, there was also a significant difference in score among the three diagnostic tools at p value < 0.0001 (F value = 51.06). For paired comparisons, there was no significant difference between diagnostic accuracy for video clips and SelectStitch at p value = 0.9391; there was a significant difference between video and Stitch (Stitch scored less) at p value < 0.0001; and there was a significant difference between SelectStitch and the Stitch (Stitch scored less) at p value < 0.0001 (see Table 2).

In summary, for both Score-1 and Score-2, diagnostic accuracy was equivalent for SelectStitch composite images and video clips, both of which were better than for Stitch images.

Question (c): How did level of confidence of ENTs for each diagnostic tool relate to their diagnostic ability?

Associations between confidence level and Score-1 were examined for each diagnostic tool. For Stitch, this association was not significant (t value = -2.4) after
Bonferroni correction (p value = 0.0168; α used for Bonferroni correction = 0.0167). For video clips, this association was positive and significant (t value = 3 and p value = 0.0027). Finally, for SelectStitch, this association was positive and significant (t value = 2.87 and p value = 0.0041). Next, we compared the magnitude of association between confidence and Score-1 among the three diagnostic tools. Results demonstrated no significant difference in this association between video clips and SelectStitch (p value = 0.9944). In contrast, the association between confidence level and Score-1 for video clips was significantly higher than that for Stitch (p value < 0.0001), and the association for SelectStitch was significantly higher than that for Stitch (p value < 0.0001). Similar findings were demonstrated for the associations between confidence level and Score-2 of each diagnostic tool. Specifically, for Stitch, the association was non-significant (t value = -2.3) after Bonferroni correction (p value = 0.0214; α used for Bonferroni correction = 0.0167). For video clips, the association was positive and significant (t value = 3.33 and p value = 0.0009). Finally, for SelectStitch, the association was positive and significant (t value = 3.19 and p value = 0.0015). Again, the magnitude of associations of confidence level and Score-2 were compared among the three diagnostic tools. There was no significant difference between video clips and SelectStitch (p value = 0.9936). However, again, the association was significantly higher for video clips than for Stitch (p value < 0.0001), and the association was significantly higher for SelectStitch than for Stitch (p value < 0.0001). In summary, the associations between diagnostic accuracy (using both Score-1 and Score-2) and confidence level were significant only for SelectStitch and video clips, and these associations were of similar magnitude.
**Question (d): What was the inter-reader variability of ENTs with each diagnostic tool?**

Kendall’s concordance was used for assessing inter-reader agreement. For Score-1, concordance was 0.5778 for Stitch, 0.4096 for video clips, and 0.4779 for SelectStitch. For Score-2, concordance was 0.584 for Stitch, 0.4312 for video clips, and 0.3529 for SelectStitch. These Kendall’s concordance values are all moderate in magnitude.

**Discussion**

Telemedicine approaches have recently been highlighted during the COVID-19 pandemic. Even before then, telemedicine started to gain increasing attention in Otolaryngology24–26. Otoscopy is well-suited to the telemedicine approach27,28, as long as a sufficient image of the TM can be obtained. One way to optimize a sufficient image is to collect a short video clip of the examination. However, this results in a digital file that is relatively large, posing a barrier to both storage and transfer, especially in remote settings27,29. We hypothesized that computer-assisted creation of a composite image would maintain equivalent diagnostic utility and physician confidence during diagnosis.

Results demonstrated that the accuracies of ENTs in making diagnoses from SelectStitch images were equivalent to those made when reviewing the full videos, regardless of how diagnostic accuracy was determined (the stringent Score-1 vs the lenient Score-2). The overall average accuracies of ENTs (specifically for Score-2) were from 57 to 74%, a range that is similar to our previous study8. However, diagnostic accuracy depended largely on the type of pathology. For example, experts were 84% accurate in diagnosing tympanosclerosis, which has some distinguishing features (i.e., discrete areas of white plaque). In contrast, accuracy was lowest for the diagnosis of TM retraction, which
can be a fairly subtle finding. Nonetheless, the most important finding of this study was that there were no significant differences in diagnostic accuracy between SelectStitch composite images and the full video clips. In contrast, Stitch composite images, which were constructed using all available frames of a given video, led to much poorer diagnostic accuracy than either SelectStitch or full video clips. This is a highly significant finding, because it suggests that single SelectStitch images provide details that are of equal diagnostic value to full video clips for expert reviewers.

Additionally, diagnostic accuracy and diagnostic confidence level were associated for both the SelectStitch and full video clips, while no association was found for Stitch images, which is interesting in light of previous work that demonstrated an overall weak relationship between diagnostic accuracy and confidence in ear experts. This finding is important, because treatment decisions are often impacted by level of diagnostic confidence of the clinician. For example, a physician may need to feel confident of providing a diagnosis of “normal” in order to choose not to prescribe antibiotics for a patient presenting with otalgia. Moreover, inter-reader agreement in this study was generally only moderate in magnitude, providing further motivation for the need to develop methods to improve the objectivity of making ear diagnoses.

This study has several limitations. First, only a subset of pathologies was included in the survey, while several important ear pathologies were excluded, such as acute otitis media. This was a result of small numbers of videos of some pathologies in our current database. Also, only videos of relatively high quality/lighting were included. Another limitation is that within-reader agreement was not evaluated. Lastly, each reader used his/her own computer monitor to evaluate the images and videos. Those monitors were
likely of different makes, models, and resolutions. All of these factors could contribute to differences in diagnostic abilities. On the other hand, our approach was likely ecologically valid; in various telemedicine settings, a variety of monitors will be used.

Although the emphasis of this study was to provide support for the value of SelectStitch composite images in potential telemedicine settings, there are other scenarios for which an otoscopic composite image is likely preferable over a video clip. For example, current electronic medical record systems are more amenable to inclusion of photodocumentation in patient charts, as compared with video examinations. Additionally, the ability to show a patient or parent a simple composite image of a TM would improve counseling, such as in providing visual confirmation of a normal ear in a child with otalgia, which may help decrease over-prescription of oral antibiotics.

Conclusion

Results of this study demonstrated that computer-aided SelectStitch composite images provide equivalent visual information as digital otoscopic video clips for ear experts to make diagnoses of different types of pathologies. Diagnostic accuracy was also found to be associated with diagnostic confidence level, and inter-reader agreement was moderate. Future studies will be required to evaluate a more diverse set of ear pathologies, as well as using videos collected under less ideal focus and lighting conditions.
Acknowledgments

The authors would like to thank Emily Luo and Benjamin Liu for curating videos and online surveys for this study.
References


15. Daniel Chaves Viquez K, Arandjelovic O, Blaikie A, Ae Hwang I. Synthesising wider field images from narrow-field retinal video acquired using a low-cost direct ophthalmoscope (Arclight) attached to a smartphone. Paper presented at:


Table 1. Proportion of correct diagnosis in percentages for each diagnostic category for each ENT physician (I through V) using Score-1 (S1) and Score-2 (S2) (%) using SelectStitch.

<table>
<thead>
<tr>
<th>Diagnostic Categories</th>
<th>ENT-I S1</th>
<th>ENT-II S1</th>
<th>ENT-III S1</th>
<th>ENT-IV S1</th>
<th>ENT-V S1</th>
<th>mean (SD)</th>
<th>ENT-I S2</th>
<th>ENT-II S2</th>
<th>ENT-III S2</th>
<th>ENT-IV S2</th>
<th>ENT-V S2</th>
<th>mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>55</td>
<td>55</td>
<td>65</td>
<td>85</td>
<td>55</td>
<td>63 (13)</td>
<td>65</td>
<td>65</td>
<td>74</td>
<td>56</td>
<td>65</td>
<td>65 (13)</td>
</tr>
<tr>
<td>Effusion</td>
<td>78</td>
<td>90</td>
<td>64</td>
<td>16</td>
<td>25</td>
<td>47 (26)</td>
<td>63</td>
<td>63</td>
<td>52</td>
<td>48</td>
<td>70</td>
<td>56 (31)</td>
</tr>
<tr>
<td>Retraction</td>
<td>38</td>
<td>55</td>
<td>29</td>
<td>48</td>
<td>37</td>
<td>39 (7)</td>
<td>56</td>
<td>56</td>
<td>45</td>
<td>40</td>
<td>70</td>
<td>56 (12)</td>
</tr>
<tr>
<td>Tympanosclerosis</td>
<td>78</td>
<td>94</td>
<td>67</td>
<td>78</td>
<td>83</td>
<td>69 (9)</td>
<td>84</td>
<td>84</td>
<td>78</td>
<td>78</td>
<td>83</td>
<td>84 (7)</td>
</tr>
<tr>
<td>Average ( mean (SD) )</td>
<td>62</td>
<td>74</td>
<td>52</td>
<td>56</td>
<td>57</td>
<td>46</td>
<td>57</td>
<td>57</td>
<td>65</td>
<td>68</td>
<td>57</td>
<td>57 (23)</td>
</tr>
</tbody>
</table>
Table 2. Comparisons of accuracy for the different diagnostic tools for Score-1 (S1) and Score-2 (S2). V = Video; S = Stitch; SS = SelectStitch.

| Label     | Estimate S1 | Standard Error S1 | Estimate S2 | Standard Error S2 | DF  | t Value | Pr > |t| |
|-----------|-------------|------------------|-------------|------------------|-----|---------|------|---|
| V vs S    | 1.1857      | 0.1424           | 1.3279      | 0.1524           | 1160| 1163    | 8.33 | <.0001 |
| SS vs S   | 1.1901      | 0.1427           | 1.3396      | 0.1526           | 1160| 1163    | 8.34 | <.0001 |
| SS vs V   | 0.0044      | 0.1331           | 0.0117      | 0.1529           | 1160| 1163    | 0.03 | 0.9736 |
Figure Captions

Figure 1. The process of Stitch and SelectStitch. In comparison to Stitch, SelectStitch possesses a deep learning based semantic segmentation step to reduce irrelevant frames from video sequences as described in 23. These excluded frames include parts of the video with low quality (e.g., those heavily blurred or having an excessive amount of cerumen).

Figure 2. An example question from the online diagnostic survey. The readers are asked to make a diagnosis of the disease either from the video or a composite image (either produced by Stitch or SelectStitch). Readers can pick one or more of the choices. If their diagnosis is not included in any of the categories, they can pick the Other Category and enter their choice (e.g. monomeric TM). Readers are also asked their diagnostic confidence level using the Likert Scale with 5 being “extremely confident.”

Figure 3. Summary of the rounds of the otoscope diagnosis survey for each reader (ENT-I through ENT-V). The order of the Stitch and SelectStitch composite images of the same sample were mixed in each survey. The cases from adult and pediatric patients were also mixed in each evaluation set.
Appendix

Appendix Table A. Proportion of correct diagnosis in percentages for each diagnostic category for each ENT physician (I through V) using Score-1 (S1) and Score-2 (S2) (%)

using *Stitch*.

<table>
<thead>
<tr>
<th>Diagnostic Categories</th>
<th>ENT-I S1</th>
<th>S2</th>
<th>ENT-II S1</th>
<th>S2</th>
<th>ENT-III S1</th>
<th>S2</th>
<th>ENT-IV S1</th>
<th>S2</th>
<th>ENT-V S1</th>
<th>S2</th>
<th>mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>15</td>
<td>15</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>33</td>
<td>35</td>
<td>20</td>
<td>20</td>
<td>17 (11)</td>
</tr>
<tr>
<td>Effusion</td>
<td>68</td>
<td>70</td>
<td>31</td>
<td>40</td>
<td>55</td>
<td>55</td>
<td>10</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>35 (26)</td>
</tr>
<tr>
<td>Retraction</td>
<td>39</td>
<td>45</td>
<td>38</td>
<td>45</td>
<td>23</td>
<td>25</td>
<td>36</td>
<td>40</td>
<td>20</td>
<td>35</td>
<td>31 (9)</td>
</tr>
<tr>
<td>Tympanosclerosis</td>
<td>64</td>
<td>67</td>
<td>28</td>
<td>33</td>
<td>38</td>
<td>39</td>
<td>49</td>
<td>50</td>
<td>43</td>
<td>50</td>
<td>44 (13)</td>
</tr>
<tr>
<td>Average ( mean (SD)  )</td>
<td>46 (24)</td>
<td>49 (25)</td>
<td>25 (14)</td>
<td>31 (18)</td>
<td>31 (19)</td>
<td>32 (16)</td>
<td>31 (17)</td>
<td>23 (13)</td>
<td>30 (16)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Appendix Table B. Proportion of correct diagnosis in percentages for each diagnostic category for each ENT physician (I through V) using Score-1 (S1) and Score-2 (S2) (%) using video clips.

<table>
<thead>
<tr>
<th>Diagnostic Categories</th>
<th>ENT-I</th>
<th>ENT-II</th>
<th>ENT-III</th>
<th>ENT-IV</th>
<th>ENT-V</th>
<th>mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>60</td>
<td>55</td>
<td>70</td>
<td>93</td>
<td>29</td>
<td>61 (23)</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>55</td>
<td>70</td>
<td>95</td>
<td>30</td>
<td>62 (24)</td>
</tr>
<tr>
<td>Effusion</td>
<td>84</td>
<td>68</td>
<td>70</td>
<td>23</td>
<td>16</td>
<td>52 (31)</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>85</td>
<td>75</td>
<td>25</td>
<td>30</td>
<td>61 (31)</td>
</tr>
<tr>
<td>Retraction</td>
<td>48</td>
<td>53</td>
<td>70</td>
<td>59</td>
<td>49</td>
<td>49 (9)</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>70</td>
<td>45</td>
<td>60</td>
<td>70</td>
<td>60 (11)</td>
</tr>
<tr>
<td>Tympanosclerosis</td>
<td>83</td>
<td>65</td>
<td>72</td>
<td>80</td>
<td>76</td>
<td>75 (7)</td>
</tr>
<tr>
<td></td>
<td>89</td>
<td>78</td>
<td>48</td>
<td>83</td>
<td>94</td>
<td>78 (18)</td>
</tr>
<tr>
<td>Average (mean (SD))</td>
<td>68</td>
<td>60</td>
<td>72</td>
<td>64</td>
<td>42</td>
<td>68 (18)</td>
</tr>
<tr>
<td></td>
<td>(18)</td>
<td>(7)</td>
<td>(13)</td>
<td>(31)</td>
<td>(26)</td>
<td>(19)</td>
</tr>
</tbody>
</table>

Page 22 of 25
Figure 1. The process of Stitch and SelectStitch. In comparison to Stitch, SelectStitch possesses a deep learning based semantic segmentation step to reduce irrelevant frames from video sequences as described in 23. These excluded frames include parts of the video with low quality (e.g., those heavily blurred or having an excessive amount of cerumen).

368x225mm (96 x 96 DPI)
Figure 2. An example question from the online diagnostic survey. The readers are asked to make a diagnosis of the disease either from the video or a composite image (either produced by Stitch or SelectStitch). Readers can pick one or more of the choices. If their diagnosis is not included in any of the categories, they can pick the Other Category and enter their choice (e.g., monomeric TM). Readers are also asked their diagnostic confidence level using the Likert Scale with 5 being “extremely confident.”

185x210mm (96 x 96 DPI)
Figure 3. Summary of the rounds of the otoscope diagnosis survey for each reader (ENT-I through ENT-V). The order of the Stitch and SelectStitch composite images of the same sample were mixed in each survey. The cases from adult and pediatric patients were also mixed in each evaluation set.

406x177mm (96 x 96 DPI)