Department of Public Health and Clinical Medicine, 2025

Ear and hearing diagnostics in low-income settings

Prospects for new eHealth supported solutions

Josefin Sandström

UMEÅ UNIVERSITY

ISBN: 978-91-8070-843-2 (print)
ISBN: 978-91-8070-844-9 (pdf)
ISSN: 0346-6612
Umeå University Medical Dissertations New Series no. 2393
Cover painting and illustrations: Johanna Eliasson
Electronic version available at: http://umu.diva-portal.org/
Printed by: Scandinavian Print Group, Skarpnäck, 2025

Nog finns det mål och mening i vår färd –
men det är vägen, som är mödan värd.

Det bästa målet är en nattlång rast,
där elden tänds och brödet bryts i hast.

På ställen där man sover blott en gång,
blir sömnen trygg och drömmen full av sång.

Nog finns det mål och mening i vår färd –

men det är vägen, som är mödan värd.

Det bästa målet är en nattlång rast,

där elden tänds och brödet bryts i hast.

På ställen där man sover blott en gång,

blir sömnen trygg och drömmen full av sång.

Innehåll

Abstract

Millions of people suffer from hearing loss globally. Otitis media and age-related hearing loss are the most common causes of hearing loss. Besides hearing loss, otitis media can lead to severe infections in the balance and hearing organs as well as the brain with serious complications. Long term consequences of both otitis media and hearing loss are preventable, but for prevention, available and reliable diagnostics is essential. South Africa is one of many countries where ear disease and hearing loss are prevalent and the studies in this thesis are conducted in low-income communities in the City of Tshwane in South Africa. The aim was to validate two new diagnostic tools based on eHealth for ear disease and hearing loss and to explore the prerequisites to implement the methods at clinics in the low-income communities.

The thesis includes two studies of smartphone-based audiometry and one study of image-based otitis media diagnostics using artificial intelligence. To explore the prerequisites for implementing these methods, the thesis also includes one study in which primary healthcare professionals working in low-income settings were interviewed about their perspectives of ear and hearing health today and for the future.

In study I and II, the accuracy and reliability of smartphone audiometry was investigated. The thresholds from the smartphone hearing tests were compared to the thresholds from conventional audiometry. When the test was conducted in an operator-controlled manner (study I), the mean difference between methods was in a soundproof booth 1.6 decibel (SD 9.9) and in a noisy primary healthcare environment -1.0 decibel (SD 7.1) (smartphone thresholds subtracted from conventional). When the hearing test was updated to a self-test (study II) and studied in a noisy environment, the mean difference compared to conventional audiometry was -2.2 decibel (SD 10.1) (smartphone thresholds subtracted from conventional). There were no significant differences between repeated measurements with the smartphone test, and smartphone audiometry was therefore considered reliable.

The third study compared the diagnosis from a convolutional neural network (one kind of artificial intelligence) to the reference diagnosis from an expert panel using otoscopic images of tympanic membranes. Accuracy was calculated and how modifications of data affected it (addition of more images, augmentation and normalisation). The accuracy was ≥88 % for all instances tested when the network was used

to discern between normal ears, obstructing wax and pathological ears. The modifications of data had no apparent effect on the accuracy.

The interview study (study IV) is ongoing, and the results are therefore preliminary. Professionals from public primary healthcare clinics were interviewed about the prerequisite for ear and hearing healthcare today, their knowledge of and attitude towards artificial intelligence and the prerequisite to implement tools like the ones proposed in study I-III at their clinics. Some important components seem to be present, although there are several aspects necessary to consider before the tools can be incorporated in the everyday work. An openness to the technique and willingness to learn and being involved was considered beneficial together with critical thinking towards possible pitfalls. It was however seen that maintenance of facilities and equipment and basic medical treatment were at instances missing, which must be prioritised first.

Overall, the studies in this thesis showed that tools supported by eHealth for ear and hearing care can contribute to accurate and accessible ear and hearing diagnostics in low-income settings. They can thereby offer possibilities to prevent long-term consequences from ear disease and hearing loss, but the unique conditions at the clinics must be considered. Future studies of AI supported tools should be performed in clinical settings.

Abbreviations

AI	Artificial intelligence
AOM	Acute otitis media
CI	Confidence interval
CNN	Convolutional neural network
CSOM	Chronic suppurative otitis media
dB	Decibel
Hz	Hertz
mHealth	Mobile health
OME	Otitis media with effusion
SD	Standard deviation
TPR	True positive rate

Populärvetenskaplig sammanfattning

Hörselnedsättning drabbar miljontals människor globalt och kan orsakas av inflammation eller infektion i mellanörat (mediaotit). Förutom hörselnedsättning kan mediaotit leda till allvarliga infektioner i balans- och hörselorganen och i hjärnan. Långvariga konsekvenser av både mediaotit och hörselnedsättning går att förebygga, men då behövs tillgänglig och tillförlitlig diagnostik. Sydafrika är ett av många länder där förekomsten av öronsjukdom och hörselnedsättning är hög. Studierna i denna avhandling är genomförda i låginkomstområden i Sydafrika, i City of Tshwane. Syftet med studierna är att validera två nya metoder för att med eHälsa diagnostisera öronsjukdom och hörselnedsättning och också att undersöka förutsättningarna för att implementera metoderna i låginkomstområden. Avhandlingen omfattar dels två studier om smartphone-baserad hörselmätning och en studie om bildbaserad diagnostik av mediaotit med hjälp av artificiell intelligens. Avhandlingen omfattar också en studie där personal från primärvårdskliniker i låginkomstområden intervjuas om förutsättningarna för att implementera de nya diagnostiska metoderna i sitt kliniska arbete.

I de två första studierna undersöktes träffsäkerheten för ett smartphonebaserat hörseltest. Hörselresultaten från smartphone-testet jämfördes mot konventionell hörselmätning. Studierna visar att smartphoneundersökningen gav kliniskt acceptabla resultat både då en undersökare testade försökspersonerna manuellt och även då undersökningen utfördes som ett självtest. Hörseltestet är studerat med likvärdiga resultat både i ett ljudisolerat rum och i en miljö med mycket omgivningsljud.

I den tredje studien jämfördes örondiagnosen ställd av en typ av artificiell intelligens som används vid bildigenkänning, ett s.k. faltningsnätverk, mot en referensdiagnos ställd av en expertpanel bestående av en grupp erfarna öronläkare. Som grund för diagnos användes bilder av trumhinnor. I studien analyserades också hur modifiering av data (tillägg av fler bilder och s.k. normalisering och augmentering) påverkade resultaten. Träffsäkerheten för nätverket var hög, ≥88 % under alla testade omständigheter då nätverket skulle skilja på normala trumhinnor, sjuka trumhinnor och skymmande vax. De modifieringar av data som användes gav varken en tydlig förbättring eller försämring av träffsäkerheten.

Intervjustudien, som utgör den sista delen i avhandlingen är pågående och resultaten därför preliminära. Personal från offentligt drivna primärvårdskliniker i Sydafrika har intervjuats om förutsättningarna att bedriva öron- och hörseldiagnostik idag, om sin kunskap och attityd till artificiell intelligens och om förutsättningarna att använda redskap baserade på artificiell intelligens på klinikerna. Det framkommer att det finns möjligheter för de nya metoderna att göra nytta, men att det också finns många aspekter att beakta innan de kan tas i bruk i den kliniska verksamheten.

Sammantaget visar studierna i denna avhandling att det finns förutsättningar att med nya diagnostiska metoder baserade på eHälsa diagnostisera öronsjukdom och hörselnedsättning i låginkomstområden. Om metoderna kan implementeras kan de bidra till ökad tillgång till träffsäker diagnostik och en minskning av de långvariga konsekvenserna av dessa tillstånd.

List of scientific papers

I Sandström J, Swanepoel DW, Carel Myburgh H, Laurent C. Smartphone threshold audiometry in underserved primary health-care contexts. International Journal of Audiology. 2016;55(4):232–8.

II Sandström J, Swanepoel D, Laurent C, Umefjord G, Lundberg T. Accuracy and Reliability of Smartphone Self-Test Audiometry in Community Clinics in Low Income Settings: A Comparative Study. Annals of otology, rhinology & laryngology. 2020;129(6).

III Sandström J, Myburgh H, Laurent C, Swanepoel DW, Lundberg T. A Machine Learning Approach to Screen for Otitis Media Using Digital Otoscope Images Labelled by an Expert Panel. Diagnostics (Basel). 2022;12(6).

IV Sandström J, Ngcobo S, Andersson J, Lundberg T. Attitudes towards artificial intelligence among healthcare professionals in unprivileged primary healthcare clinics in South Africa, with a focus on ear and hearing diagnostics – an interview study. Preliminary draft

Background

Ear and hearing epidemiology

Ear and hearing health are closely intertwined. Ear disease can, if left untreated, progress to a permanent hearing loss and hearing loss can in turn be the first sign of disease in the ear. Besides hearing loss, ear disease and specifically infections can have serious complications including labyrinthitis, meningitis and brain abscess that can cause permanent sequelae or death [1-3]. Hearing loss is called a silent epidemic since it is affecting about 20 % of our global population with an increasing prevalence [4]. In the next 30 years, it is expected that the number of people living with hearing loss will rise more than 50 % and affect 2.45 billion people because of aging populations [4]. The body of evidence mapping out the consequences of hearing loss is also growing. It is well-known that hearing loss is detrimental to academic performance and opportunities to earn one's livelihood [5,6]. Furthermore, people with hearing loss have lower social and emotional wellbeing as compared with people with normal hearing. Among the more recent additions to this list is the established link between hearing loss and dementia [7,8].

The leading cause of hearing loss is the age-related type (presbyacusis). Second is inflammation in the middle ear (otitis media) [9]. Otitis media is an umbrella term for several conditions including acute otitis media (AOM), otitis media with effusion (OME) and chronic suppurative otitis media (CSOM) [3]. There are means to treat otitis media and thereby prevent complications, but for treatment, an accurate diagnosis is necessary. In addition to omitted otitis media treatment, and a problem to highlight, is over-diagnosis of AOM, leading to use of antibiotics when not indicated [3,10]. That practice may aggravate antibiotic resistance, which is a major health concern globally, jeopardising the overall health, and routine healthcare procedures [11,12]. AOM is however one of the most common infections in children globally which under certain circumstances needs antibiotic treatment [3].

The burden of different ear and hearing conditions is unevenly distributed globally. CSOM is more common in low- and middle-income countries and hearing loss due to otitis media is relatively more frequent in these regions as compared to high-income countries [3,13]. Furthermore, the incidence of congenital permanent hearing loss is more common in

low-income countries explained to a large extent by insufficient maternal and neonatal care [14]. In high-income countries on the other hand, presbyacusis is the predominant cause of hearing loss [4]. Other contributors to hearing loss are noise exposure (occupational and recreational), tobacco smoking and the use of ototoxic drugs [15-17]. All these risk factors are therefore important targets in the prevention of hearing loss. Moreover, access to healthcare is essential.

Hearing and the middle ear

Hearing and hearing anatomy

The auditory function is an intricate process requiring intact organs for transmission of soundwaves and neural signals, and cognitive capacity to handle these signals. The soundwaves reach the outer ear and are transferred through the external ear canal to the tympanic membrane (Figure 1). The membrane vibrates in accordance with the soundwaves and conducts the vibrations to the ossicular chain in the middle ear. The end of the ossicular chain (the footplate of the stapes) connects to the cochlea, which is the auditory organ of the inner ear. In the cochlea, the information is transformed into neural signals that travel to the brain regions involved via the auditory nerve and central auditory pathways. Damage to any part of this system can affect the hearing in different ways, depending on what part that is malfunctioning. Reduction in brain function, temporary or permanent, can therefore also hamper the hearing

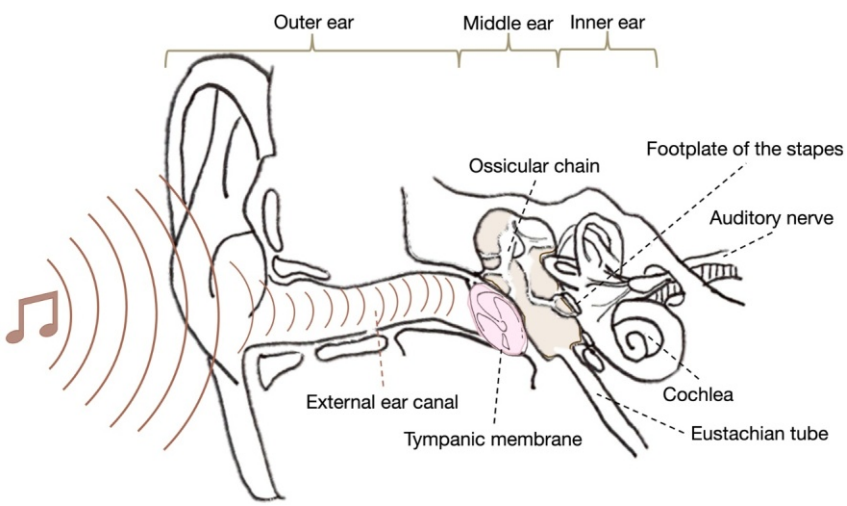


Figure 1. Anatomy of the ear and hearing organs and the path for sound transmission.

ability. A broad subdivision of hearing loss can be made into conductive and sensorineural hearing loss. The first is caused by damage of the conductive components, from the ear canal to the end of the ossicular chain. Sensorineural hearing loss is due to impaired neuro-signalling, from the cochlea or auditory nerve [6].

Hearing is reported as hearing thresholds in the decibel scale (dB). The range for what humans are able to hear is wide, and the dB scale is therefore logarithmic. The hearing threshold is defined as the level where a tone is heard 50 % of presentations at that specific level [18]. The higher the hearing threshold is, the poorer is the hearing. Usually, steps of 5 dB are used to determine the thresholds [18]. Hearing thresholds are recorded for multiple frequencies and can be visualised as an audiogram (Figure 2). It is common to measure at 0.5, 1.0, 2.0, 4.0 and 8.0 kHz since these frequencies encompass many of the everyday sounds. In certain circumstances other frequencies can be of interest, for example when monitoring the hearing during treatment with ototoxic drugs that are known to affect higher frequencies [19]. The hearing capability can also be reported as the pure tone average for the frequencies most important in apprehension of normal speech (0.5, 1.0, 2.0 and 4.0 kHz). The classification of hearing loss differs slightly in different standards and has been subject to change. The World Health Organization uses the following levels (pure tone average in the better hearing ear): mild hearing loss 20-<35 dB, moderate hearing loss 35-<50 dB, moderately severe hearing loss 50-<65 dB, severe hearing loss 65-<80 dB, profound hearing loss 80-<95 dB and complete hearing loss/ deafness ≥95 dB [6]. The benefit from a hearing aid is individual but is seen already from mild hearing loss [20].

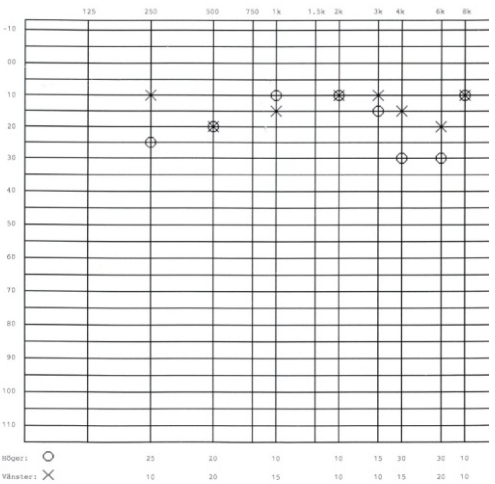


Figure 2. Result from audiometry, presented as an audiogram.

Hearing assessments

Gold standard for hearing assessment is pure tone threshold audiometry [6]. It is traditionally performed in a soundproof booth and run by an audiologist. Both air-conducted and bone-conducted thresholds can be established. Bone conduction is used to distinguish between conductive and sensorineural hearing loss, but it was not part of the procedures in the audiometric studies in this thesis and will not be further described. The gold standard method requires both specific audiological equipment and trained audiologists, which is lacking in major parts of the world [21,22]. There are also several simpler methods for assessing hearing that can be suitable in specific situations and in underserved settings. Among these methods are the free field voice test, tests using tuning forks (Weber's and Rinne's test) [23] as well as forms for self-reported hearing loss [24].

There is inherent variation in a person's hearing that does not reflect a true change in the auditory system per se. This is due to the psychoacoustic nature of audiometry [25]. This factor is difficult to predict but implies that the hearing threshold can be lower at a first assessment and higher the next, or the opposite, without a physical change in the system. This is because the hearing is not dependent only on the auditory organs, but also on what happens in the processing phase where there is competition with other brain functions. The patient's temporary mode, the ability to keep up concentration on the specific day for the test and the surroundings in the test situation can hence affect the hearing thresholds. It is also exhaustive to sharpen the hearing for a long time, leading to a decline if a prolonged test protocol is used. Furthermore, the brain itself can have the impression of sound when there are no actual outer stimulus and people suffering from tinnitus can struggle to hear even though there is not always a clinical hearing loss [6]. From all these factors follow that the result of a hearing test is not expected to be identical if repeated, even when everything that can be controlled is controlled.

The middle ear anatomy

The middle ear is the part of the ear between the tympanic membrane and the inner ear (Figure 1). The connection between the outer ear and the tympanic membrane is the external ear canal. There is also a connection between the middle ear and the nasopharynx (located behind the nasal cavity) called the Eustachian tube, a narrow canal that

evacuates secretion and equalises the pressure in the middle ear to the atmospheric pressure. Assessment of the middle ear can be done by *inspection* and by *physiological tests*. Several conditions in the middle ear and also in the external ear canal can be diagnosed this way, and those that will be mentioned in this thesis, and that are among the most common, are AOM, OME, CSOM, obstructing ear wax, otitis externa and foreign bodies (Figure 3).

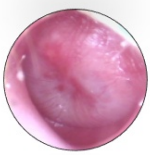
Middle ear assessments

Inspection of the middle ear is partwise an inspection of the tympanic membrane, since it covers the lateral entrance to the middle ear, but structures in the middle ear can often also be visualised through it (Figure 4). For inspection, there are several tools to use, the most basic are head lamps and otoscopes, and more costly are otomicroscopes and digital video-otoscopes connected to laptops, smartphones or tablets. The appearance and position of the tympanic membrane should be described, and whether there is a perforation in it or fluid behind it. A healthy tympanic membrane is intact, more or less translucent so that some of the middle ear structures can be seen through it, and it has a pearly white or grey appearance. It should be neither bulging nor retracted and no fluid should be seen behind it. When inspecting the ear, the external ear canal is also examined.

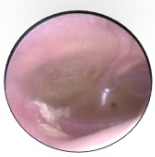
The physiological tests of the middle ear serve, as a complement to the inspection, to decide whether there is fluid in the middle ear. Fluid can be purulent or clear, the first a feature of infection (AOM) and the second a feature of OME, an inflammatory condition [3]. The tests can also assist in diagnosing a perforation, as in CSOM, if it is not possible to visualise it. The most common methods to evaluate the presence of middle ear effusion are pneumatic otoscopy and tympanometry. With pneumatic otoscopy, an insufflator bulb is connected to the otoscope and a brief portion of air is applied to or removed from the external ear canal to alter between a negative and positive pressure. The examiner observes if there is mobility of the tympanic membrane when the pressure is changed, as in a healthy ear [26]. In tympanometry, a sealing ear plug is positioned in the opening of the external ear canal. It is connected to the tympanometer, which controls and alters the pressure in the ear canal and emits a continuous sound. Part of the sound is transferred to the middle ear, and parts are reflected. By measuring the reflected sound, the mobility of the tympanic membrane and the pressure in the middle ear can indirectly be decided. The volume of the external ear canal is also

Diagnoses and management

Normal tympanic membrane The healthy tympanic membrane is translucent, of normal thickness, pearly grey in appearance, and neither retracted nor bulging. Some of the structures in the middle ear can be seen through it and be used as landmarks in the assessment. The membrane is mobile when assessed with pneumatic otoscopy and tympanometry.



Acute otitis media An infection in the middle ear most common in children <5 years of age. The diagnosis requires both acute symptoms and otoscopic findings. Symptoms can be earache and fever and when inspected, the tympanic membrane is opaque and bulging and the landmarks in the middle ear are not seen. Purulent fluid can be seen behind the tympanic membrane, or in the external ear canal if the membrane is perforated. There are acute, as well as chronic complications of AOM, ranging from life threatening meningitis to life long hearing impairment. There are cases of AOM which should be treated with antibiotics, but guidelines between countries differs on when to do that.

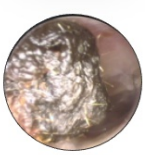


Otitis media with effusion An inflammatory condition with non-purulent fluid in the middle ear. It can occur in the mending phase of AOM or when dysfunction or blockage of the Eustachian tube. Most often the condition is without symptoms and self-resolving, but the fluid might block the transmission of soundwaves and affects the hearing. The reason for OME should be determined, and in children with persistent fluid and affected hearing there might be indication for placement of ventilation tubes.

Chronic suppurative otitis media Unresolving inflammation of the middle ear, with a chronic perforation, and commonly a complication of AOM or OME. The condition affects the hearing. Several measurements for treatment are usually needed, among them local or systemic antibiotics and surgery. CSOM is said to be inactive when it is not draining any more, but still with low chance to spontaneously resolve.

Foreign body All that does not emanate from the ear itself is referred to as foreign bodies. It can be beads that children put into the canal, bugs that decide to creep into it or ventilation tubes from surgery that have been discarded from the tympanic membrane.

Otitis externa This is an inflammation of the skin that covers the external ear canal. The first symptoms are itchiness and sometimes light discharge. The canal becomes swollen and can be painful if pronounced. There are both infectious and non-infectious causes to inflammation of the ear canal. Topical ointments are used for treatment.



Wax Production of wax is a normal activity in the external ear canal. The canal is a self-cleaning system, and the wax usually follow the discarded epithelial cells outwards. If the migration becomes disturbed, wax is not cleared properly. When wax is accumulating, it does not only cause problems to the examiner who can no longer see the tympanic membrane, but it can also cause hearing loss if there is a complete impaction. In such cases, syringing or other means to get it out can be necessary.

Figure 3. Characteristics and management of the most common ear conditions. AOM = acute otitis media, OME = otitis media with effusion, CSOM = chronic suppurative otitis media.

estimated by the tympanometer, which is useful in diagnostics of perforations and assessment of ventilation tube functionality [27].

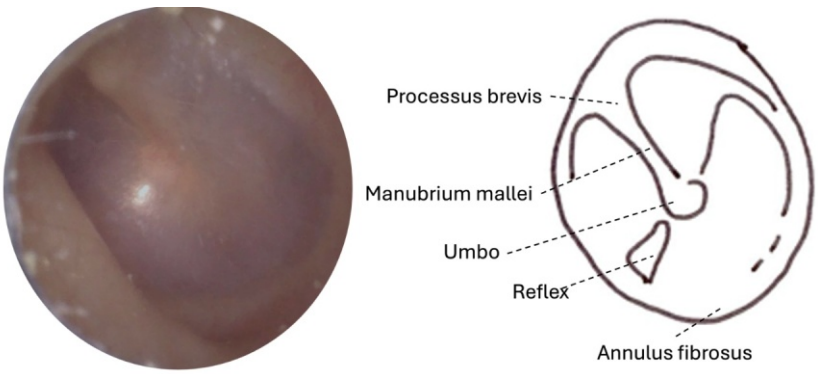


Figure 4. Normal tympanic membrane and structures to look for and assess.

Access to reliable diagnostics

Diagnostic accuracy of middle ear disease

Diagnostics of middle ear disease is difficult. As an example, general practitioners in one study diagnosed 39 %– 46 % of cases correctly when having 4 diagnoses to choose from [28]. Otolaryngologists' accuracy on the same cases was 61 %–74 % [10,28]. Of ears diagnosed as AOM by a general practitioner in another study, 78 % was given the same diagnosis by an otolaryngologist [29]. Among the more common conditions, OME seems the most difficult to diagnose. General practitioners were correct in 53 % of cases when differentiating ears with middle ear effusion from those without [30]. With patients under anaesthesia, otolaryngologists have a sensitivity of 88 %–96 % for the same diagnosis when using high magnification otomicroscopes [31,32]. However, studies of diagnostic accuracy differ widely in methodology. It makes it difficult to compare results, but still possible to assume that middle ear diagnostics is difficult.

Access to ear and hearing healthcare

Ear and hearing healthcare is often part of the primary healthcare assignment and provided in the first line by general practitioners and nurses. In the second line, otolaryngologists and audiologists are the

caregivers. For both these levels of care, there is a lack of professionals. This is true for both low- and high- income countries, even though the magnitude differs [21,22,33,34].

eHealth solutions

Concepts of eHealth

The digital development has been emphasised as an enabler for improved health in countries with poor access to healthcare [35]. Utilising digital technologies for medical purposes is known as eHealth. The applications are multifarious and include on-site as well as tele-medical solutions. Tele-medicine is the concept of providing healthcare at a distance, which is widely employed in many disciplines. Mobile health (mHealth) involves using mobile wireless devices, such as smartphones and tablets, for medical purposes. The term digital health is suggested as an overarching conceptual term encompassing all these approaches and even so the rising field of artificial intelligence (AI) in healthcare [35]. Automation of assessments is another feature discussed in this context, to simplify procedures. It opens up for less trained personnel to conduct examinations otherwise requiring qualified staff, a concept known as task-shifting [36]. The examination procedure should be simple enough to, after minimal training, be safely operated with minimal patient risks, e.g. giving instructions and applying devices. Some methods based on eHealth, mHealth, tele-medicine and automation of procedures are being implemented within ear and hearing healthcare [37-39].

Artificial intelligence

The concept of AI has existed since the 1950s but has had an exponential spread in our societies more lately. Healthcare is no exception. It is a field of computer science dealing with tasks typically requiring human intelligence. There is no single definition, but the concept includes automated systems learning from data to solve a task. It also includes correction of itself and improvement when receiving some kind of feedback and has been proposed as a measure to improve access to and accuracy in healthcare [40]. There are a few sub-concepts of AI that will be used in this thesis, and they are therefore introduced here below.

• **Machine learning** is one kind of AI with the ability to transform input data into an output by means of data-driven rules stemming from the multiple input examples fed into the system [41]. Machine learning

repeatedly recognises patterns in data and learn from these patterns [40]. The original technique was limited since humans were required to program how raw data should be transformed into useful information in the system (feature extraction) [42].

• **Deep learning** is a kind of machine learning with the advantage that feature extraction does not have to be designed by humans, but instead is driven by the data itself. It enables increased capacity to handle and compute data [42]. The implications in healthcare are broad, with examples such as detection of malignant skin lesions and keeping health records [41]. The drawback is that it is unknown how the system makes its conclusions, which is termed the black box [43].

• **Convolutional neural networks (CNN)** are deep learning networks with a structure making them easy to train in a generalised way. They are suitable for object classification and detection tasks in images [42]. The networks have a central role in computer vision, the domain in which AI is used for interpretation of images or videos [41].

• **Supervised learning** is machine learning based on labelled data and the most common form of machine learning [42]. It means that the input training data is assigned to a specific category by a human examiner, e.g. a picture of the tympanic membrane assigned with the diagnosis of AOM. The model can then be used to predict the diagnoses of new images. It takes a lot of work to obtain the amount of labelled data necessary for supervised learning. Collecting and labelling small datasets for specific tasks is often feasible, but results in algorithms with poor performance on new data [41].

• Algorithm training and modifying training data — augmentation and normalisation

In the evaluation of machine learning algorithms for image classification, testing should be done on images that have not been used to train the model [44]. A proportion of images in a dataset is therefore left outside the training of the system. The proportions used are often 75-80 % of data for training and 20-25 % for testing. The algorithms perform better the more training data there is, and as mentioned in the section above, this is often a limitation. A method used to amplify training data without actually acquiring more real samples is *image augmentation* [45]. Augmentation can be done by e.g. rotate, shift or flip an image to create a new one. The new image will be similar to the original one but not the same and thereby unique and available for algorithm training. Since

these images are similar, there is however a limitation in how far this method can be upscaled.

Another property of machine learning for classification is how the performance is affected when categories of training data are imbalanced (e.g. more images of AOM than CSOM). The algorithms then tend to perform imbalanced too, favouring categories comprising more training data at the expense of the categories less prevalent. One approach to handle this bias is to reduce the number of images in the abundant categories, a concept called under-sampling, which is a type of normalisation [46]. The modifications of augmentation and normalisation were used in study III in this thesis.

Scope of the thesis

Smartphone audiometry

Audiometry performed by audiologists in soundproof booths are assessments requiring infrastructure and human resources that are expensive and limited [21,22]. There are methods for hearing assessments without soundproof booths, instead employing earcups or earphones proven to minimise the impact from ambient noise [39,47]. By utilising them, hearing tests can be conducted on customary smartphones furbished with a suitable software. By the time data was collected for the studies in this thesis (2014 and 2016), the first studies had been undertaken both on how mHealth solutions could be used for hearing screening (pass or fail on a predefined screening level) and also to decide the exact hearing level [48-57]. Among the challenges were that many of them were using the earphones provided by the manufacturer, and there were no means to calibrate them in a standardised way [49,51,53]. Furthermore, the majority was designed for iOS devices [49-51,53], which in general are more costly and thereby not the first choice in underserved settings. We therefore wanted to investigate smartphone audiometry for Android devices with equipment calibrated according to standards.

Image-based middle ear diagnostics using AI

Since one leg in otitis media diagnostics is the visual assessment of the tympanic membrane, and since this assessment is known to be difficult, there are reasons to assume that methods using AI proven reliable in other visual tasks could enhance the diagnostic accuracy. This thesis includes one study investigating the performance of a CNN for otitis

media diagnostics. When the study was conducted (2016), there were other researchers utilising AI for tympanic membrane assessments focusing on finding useful models. One problem with the earlies models was that the systems were fixed and could not be modified [58,59]. In another of the early studies, the model was train to decide whether there were ventilation tubes in the tympanic membrane or not [60]. It is a narrow exercise that might not have a high clinical relevance [61], especially not in low-income settings, but still important in getting to know the technique. Furthermore, the training data used in the models was of varied quality, and if the ground truth diagnosis is uncertain, it is also difficult to evaluate the outcome. Because of this, we wanted to investigate a machine learning model trained on a solid ground truth diagnosis for ear conditions of clinical value in low-income settings.

Prerequisites for eHealh solutions in low-income settings

Introduction of new technology in healthcare is more likely to be successful when the presumptive users are involved [62]. Not only information of the existing situation gives indication of potential hampers, but also what the users know and think of the new technology. Among the earliest applications of AI in healthcare was in radiology. When exploring the views of different stakeholders, radiologists have had a high representation [63-66]. A study from France from 2020 in which different stakeholders were included, 5 of 13 doctors were radiologists. The study revealed that there are many myths related to AI, and many based their knowledge from popular culture [63]. Concerns regarding the liability aspect were also raised. The views from other groups of healthcare staff have subsequently been explored. Those working in image-based domains are most represented, but primary care clinicians' views have also been investigated [67,68]. What is less represented are healthcare professionals working in underserved settings [69]. Furthermore, a majority of studies are based on online surveys [66], which constraint answers to predefined alternatives or free text fields, but with no opportunity to dive deeper if interesting themes are seen. The methods for ear and hearing diagnostics investigated in this thesis can be evaluated on performance and technical functionality, but that is not enough for a favourable use in clinical practice. Therefore, the fourth study in this thesis was designed to explore the prerequisite for introducing diagnostic solutions as the ones investigated in study I-III in the same environment where they were evaluated. In-depth interviews with professionals from primary healthcare clinics in low-income settings were conducted 2024.

Research questions

Technological progress can when deployed within healthcare improve access to and accuracy of diagnostics. In this thesis, new eHealth methods for ear and hearing diagnostics are investigated in low-income settings in South Africa. Furthermore, it is explored how the staff consider implementation of such techniques for clinical practice. Can automated smartphone audiometry with inexpensive and calibrated equipment provide accurate hearing measurements? Can AI for otitis media diagnostics provide accurate diagnoses on images when using a solid ground truth diagnosis? How can these methods be incorporated in the existing healthcare system?

Aim

Overall aim of the thesis

To validate new eHealth solutions for ear and hearing diagnostics and to investigate the prerequisite to implement the methods in clinical practice in low-income settings in South Africa.

Specific aims for the included papers

Study I
To validate the accuracy and reliability of a smartphone hearing test, operated by an examiner, in a soundproof booth and in a noisy primary healthcare environment in low-income settings.

Study II
To investigate the accuracy and reliability of the smartphone test when updated to a self-test and studied in a noisy primary healthcare environment in low-income settings.

Study III
To investigate the performance of a machine learning algorithm for diagnostics of otitis media on tympanic membrane images labelled by an expert panel and to investigate the effect of modified data.

Study IV
To explore the circumstances for ear and hearing diagnostics and the prerequisite for implementation of the proposed methods in primary healthcare clinics in low-income communities.

Materials and Methods

Overview

All data for the studies in this thesis was collected in the City of Tshwane (the metropolitan municipality around Pretoria), Gauteng province, South Africa. Data was collected 2014, 2016 and 2024.

Study I and II compare hearing thresholds from conventional audiometry to those captured by smartphone audiometry. An intra-individual, repeated-measurement method was used. In study I, smartphone audiometry was controlled by an examiner, while it in study II was conducted as a self-test. In study I, the measurements were done in two different settings. The first was a controlled setting, a soundproof booth which is the gold standard for audiometry. The second setting was not controlled, measurements took place in noisy primary healthcare clinics in underserved areas where no soundproofing was available. In study II, the setting was comparable to the uncontrolled setting in study I.

Study III is a diagnostic accuracy study in which a machine learning algorithm was used for middle ear diagnostics. Tympanic membrane images were collected using a digital video-otoscope. The reference diagnoses were assigned by an expert panel and compared to the diagnoses from a CNN.

Study IV is an ongoing qualitative study in which professionals at public primary healthcare clinics were included. Semi-structured interviews were used to explore the prerequisite for ear and hearing diagnostics at the clinics and the knowledge of, and attitudes towards AI to assist in this. For analysis of data, qualitative content analysis was used.

The smartphone hearing tests (study I and II), collection of tympanic membrane images (study III) and interviews (study IV) were performed by the author of this thesis. Conventional audiometry was done manually by registered audiologists in the controlled setting in study I. In the noisy setting (study I primary healthcare clinics and study II), conventional audiometry was automated and performed by registered audiologists or audiology students trained in this method. For overview of the methods, see Table 1.

	Design	Participants	Outcome
Study I	Intra-individual, repeated measurements	96 adults	Difference between audiological methods
Study II	Intra-individual, repeated measurements	63 adults	Difference between audiological methods
Study III	Comparison of diagnosis from artificial intelligence with reference diagnoses	190 children and adults	Diagnostic accuracy of middle ear images
Study IV	In-depth interviews	17 professionals from primary healthcare	Preliminary summary of codes from 7 participants

Table 1. Design, included participants and main outcome for study I-IV.

Environment for data collection

South Africa is by the World Bank classified as an upper middle-income country [70]. However, it is a country housing among the largest inequalities in the world of how wealth is distributed, with 10 % of the population owning 86 % of the aggregated wealth [71] and a majority of the population living in poverty [72]. Inequality in South Africa is not only a matter of wealth, but also health, and private insurances is a contributing factor. People without insurance have lower access to care [73]. There is governmentally run healthcare in South Africa, but it suffers from under-financing. Primary healthcare is for free for people not having a private health insurance. Approximately 16 % of the population can benefit from private healthcare and hence the majority rely on the under-financed public clinics [73,74].

South Africa has the highest number of people living with HIV in the world. Close to eight million are infected, constituting 13 % of the population [74] and approximately 72 % of them are treated with antiretroviral drugs [75]. HIV affects the health status overall due to susceptibility to other diseases and by the cost for HIV care. People infected with HIV are furthermore more prone to hearing loss because of opportunistic infections. Tuberculosis is one more serious and highly prevalent infection with major impact on the overall health [76]. Treatment

for both tuberculosis and HIV are associated with hearing loss, contributing to a high prevalence in South Africa [6,17,21].

Other conditions affecting the overall health in South Africa are a high prevalence of overweight (68 % of women and 23 % of men) and an increasing prevalence of diabetes, affecting approximately 6 % of the population [74]. Alcohol consumption in South Africa is among the highest in the world, which has, among other things, serious implication on maternal and children's health. Violence, trauma and injuries are also great concerns, not only affecting health, but the everyday life in many ways [74]. Society furthermore battles a high unemployment rate, approximately 30 %, and even higher among the young [74,77].

Participants

For study I, the participants in the controlled setting were recruited from individuals visiting either the Department of Speech-Language Pathology and Audiology at the University of Pretoria or the Audiology Department at Steve Biko Academic Hospital, Pretoria. They were adult patients coming for a first appointment or follow up for known hearing problems, but also some volunteers interested in having a hearing test done were included. Participants in the uncontrolled setting (Figure 5) were recruited from two different primary healthcare clinics, Stanza II and Daspoort clinics, in unprivileged areas in the City of Tshwane. Eligible in this setting were those who failed hearing screening. Hearing screening was part of the service provided at these clinics by students from the Department of Speech-Language Pathology and Audiology at

Figure 5. Waiting area next to the room for hearing assessments and busy examination room at the Stanza clinic.

the University of Pretoria, as part of their practical training. The procedure was to, following a positive screening result, continue with threshold audiometry and that was why those patients were eligible. Data collection was done in September and October 2014.

Participants for study II and III were recruited concurrently. The sites for recruitment were three different primary healthcare clinics (Stanza I, Stanza II, and Daspoort clinics) one ENT clinic (Mamelodi district hospital), and a school screening event at Mamelodi West. All participants were volunteers from the waiting area, some queueing for ear or hearing consultations but some were just accompanying others or seeking care for other reasons. All participants in study II were also enrolled in study III, but not everyone in study III was included in study II. This was because some of them did not fill the inclusion criterium of age ≥18 years for study II or that there was no equipment for hearing assessment available on the specific day. Also, the quota for statistical power for study II was eventually met and no further hearing tests were needed for study purposes. The inclusion criterion (Figure 6) of dry ears in study III was used because this was the requirement for participation in the audiometric study (study II). The exclusion criterium of unilateral hearing loss >40 dB was because this can cause overhearing. Data collection took place between March and May 2016.

Inclusion criteria
Study I: age>18 years
Study II: age ≥18 years, not actively draining ears or moisty external ear canal, complete test with reference method
Study III: age >1 year
Study IV: nurses, doctors and audiologists at the selected primary care clinics

Exclusion criteria
Study I: unilateral hearing loss >40 dB, poor concentration or cooperation, frequencies reaching the maximum smartphone hearing test output level
Study II: unilateral hearing loss >40 dB, non-compliance to index test
Study III: actively draining ears or moisty external ear canal
Study IV: professional involvement with AI

Figure 6. Inclusion and exclusion criteria for study I-IV.

For study IV, participants were healthcare professionals attending patients with ear and hearing problems in their daily work. They were either doctors or nurses at different levels of their training or audiologists. Recruitment was done by purposive sampling from five different public primary healthcare clinics (not specified because of confidentiality reasons). They were prior to data collection informed of the study via the manager at the respective clinic and at the day for the interview also by the researcher before signing the informed consent form. Data was collected in April and October 2024.

Equipment

Equipment for audiometry

The index test in study I and II was smartphone audiometry. The underlying technique had been used for hearing screening in previous studies [78], i.e. to decide whether participants could hear (yes or no) at a predefined screening level. The software was run on customary Samsung smartphones (Galaxy S3 GT-19300 for study I and Trend Neo for study II). Sennheiser headphones (HD 202 II for study I and HD 280 pro for study II) were used as transducers, nor them specifically made for audiometry, but used to reduce the cost for equipment. The same pair of smartphone and headphones was used throughout the respective data collection.

Conventional audiometry in study I was performed utilising two different audiometers, one for the controlled environment and another in the noisy environment. In the controlled setting, a clinical audiometer, GSI 61, was used in conjunction with TDH49 headphones, equipment specifically manufactured for audiometry. Both the conventional and the smartphone hearing tests were in the controlled setting conducted in soundproof booths certified for diagnostic testing. In the noisy environment in study I and II, a portable diagnostic audiometer (KUDUwave, eMoyoDotNet) was used to decide the conventional thresholds [39,79-82]. This audiometer was already employed at the clinics. It can be used outside a soundproof booth by covering insert earphones with earcups to block out environmental noise (Figure 7). There are also microphones registering the surrounding sound level and audiometry is paused if the noise rises above the permitted limit for KUDUwave testing. The portable audiometer was connected to a PC laptop where the thresholds were recorded. All equipment was calibrated prior to study

start according to ISO-standards to ensure that the output of the headphones corresponded to the recorded sound pressure level.

Figure 7. Earcups of the KUDUwave audiometer.

Equipment for image collection and CNN diagnostics

For collection of tympanic membrane images, a digital video-otoscope was used (Welch Allyn Digital Macroview). It was connected to a PC laptop where the tympanic membrane was visualised on a 15“screen. Images captured were stored on the internal memory of the laptop in the PNG format and had a resolution of 1280x1024 pixels. Different sizes of ear specula were available.

The digital video- otoscope used was from a well-established manufacturer and of high quality, but also pricy. Part of the preparations for image bank data collection was to explore a less costly alternative. Experiments were made to use the camera incorporated in a Sony XPERIA Z2 smartphone to capture images. At first, a customary ear speculum for traditional hand-held otoscopes was connected to the camera via a clip-on solution, but it was difficult to keep it fixed while manipulating the speculum in the ear canal. By 3D-printing a more firmly fitted attachment this problem could be solved. As light source, the light from the camera itself was initially used, but a too large portion of it was absorbed by the speculum. By mounting external light diodes, the tympanic membranes included in study III were also photographed with acceptable result with this own-produced, low-cost smartphone-otoscope (Figure 8). However, the images were not used for further analysis due to limited overview of the tympanic membrane.

For AI diagnostics, the CNN known as Inception v1 or GoogLeNet was used [83]. It was freely available at the Google TensorFlow platform and was used without any modifications of the architecture.

Figure 8. Smartphone otoscope prototype and a tympanic membrane image captured with it.

Equipment for interviews and analysis of interview data

The interviews were recorded on a digital voice recorder (Olympus DS-2600). They were stored on a memory card in the mp3 format. To organise data and facilitate analysis of the interviews, the MAXQDA software (version 24.8.0) was used.

Procedures

Procedures study I and II

All participants underwent two different air conducted pure tone audiometric tests, a conventional test and a smartphone test (Table 2). The tests were done in sequence with a few minutes resting time in between. For all tests, octave frequencies between 0.5 kHz and 8.0 kHz were included in the test protocol starting at 1.0 kHz, followed by 0.5 kHz, 2.0 kHz, 4.0 kHz and 8.0 kHz. A re-test at 1.0 kHz was included in the end of each session to be used for reliability calculations. In study I, 53 % of participants started with the conventional test. In study II all participants instead started with the conventional test, since it had a feature to make sure instructions were understood correctly before the real test began.

	Study I	Study I	Study II
Environment	**Soundproof booth**	**Noisy**	**Noisy**
Conventional test	**Manual standard audiometry**	**Automated KUDUwave**	**Automated KUDUwave**
Smartphone test	**Operator-controlled**	**Operator-controlled**	**Self-test**

Table 2. Environment, and method for conventional test and smartphone test in study I and II.

Conventional audiometry — performed by audiologists, as well as with the KUDUwave in automated mode — was conducted using the modified Hughson-Westlake threshold-seeking method. The smartphone test was in study I managed as an operator-controlled test, in which an examiner adjusted the sound level according to the participants' responses. In the test, a repeated tone was played for 750 ms followed by a pause of 500 ms. The sound level was adjusted by a slider on the smartphone screen (Figure 9). The participant indicated that the tone was heard by raising a hand, and when it could not be heard any longer, the hand was taken down. In a first step, the sound level was modified in 10 dB-steps to establish a crude hearing threshold. This was done by reducing the volume when the hand was in the air and turning it up when the hand

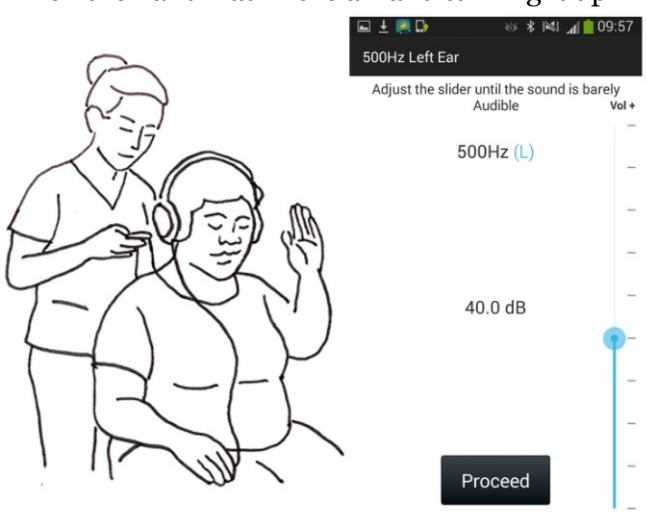


Figure 9. Arrangement and screenshot of the operator-controlled smartphone hearing test.

was taken down, repeatedly to ensure consistency. Thereafter, 5 dB-steps were used for fine-tuning, starting 5 dB above the crude threshold. The examiner stood behind the participant to minimise the risk of a visual impact (Figure 9), both by psychoacoustical reasons and by unintentional body language indications. In study II, the software had been further developed into a self-test. In the self-test, a tone was presented and the participant responded by pressing a round clear button on the smartphone screen when the tone was heard (Figure 10). This automated smartphone test was programmed to follow the ISO shortened ascending method. The thresholds from manual audiometry were written down on a results sheet directly by the audiologist. The hearing thresholds from the remaining testing were stored automatically on the respective device and from there manually transferred to the results sheet. All results were subsequently manually transferred to an Excel spreadsheet. Testing results from the other method were blinded to the examiners.

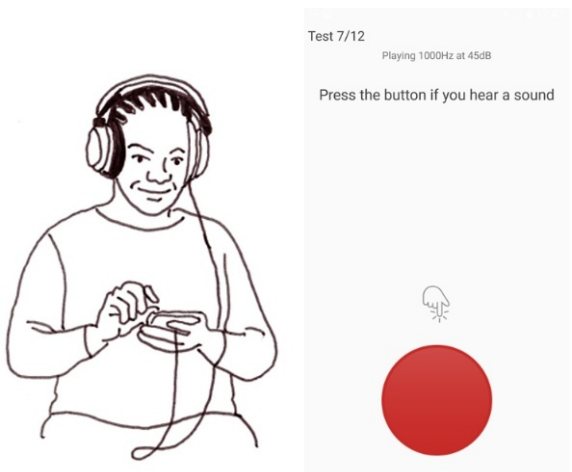


Figure 10. Arrangement and screenshot of the smartphone self-test.

In study I, the lowest stimulus for the smartphone hearing test was 5 dB for 0.5 kHz, 10 dB for 1.0 and 2.0 kHz, 0 dB for 4.0 kHz and 5 dB for 8.0 kHz. The conventional audiometer could test down to 0 dB. This induces a floor effect when a person with 0 dB hearing level is undertaking the smartphone test, which will register a threshold e.g. 10 dB at 1.0 kHz. To circumvent this, only thresholds >15 dB with conventional method was subsequently analysed. In study II, both the smartphone and the conventional hearing test were programmed to measure down to 10 dB.

Procedures study III

A series of pictures of the participants' tympanic membranes was collected, the number varied depending on how well the membranes could be visualised. The participants both ears were photographed, which generated 380 images in total (Figure 11). After data collection, the 2-5 best images from each ear were selected for inclusion in the image bank by two physicians experienced in otology. The single best image from each ear was also tagged.

To generate reference diagnoses for the collected images, an expert panel of five otologists was recruited. Two of the experts were from South Africa and three from Sweden. Their experience in middle ear assessments was at a minimum 20 years. The expert panel was instructed to choose one of the five categories: 1. Normal, 2. Obstructing wax (will be referred to as Wax), 3. Acute otitis media, 4. Otitis media with effusion and, 5. Chronic suppurative otitis media, which in the study consisted of inactive CSOM (dry chronic perforation, because draining ears were excluded). There was also an option for the panel members to pass (in the published article referred to as not possible to determine, NPD), if they from the images available could not ascribe the image with one of the five diagnoses. With this last option, they had a total of six alternatives to choose from. The panel underwent a calibration procedure on how to diagnose from defined diagnostic criteria. Twenty images were used for this, and a few queries were solved. That procedure, together with exclusion of 13 images due to missing files and discharging ears made 347 images available for further assessment (Figure 11). The panellists did their assessments individually via a Google forms document wherein the images and alternatives for diagnosis were presented. For each ear, the 2–5 selected images were displayed in sequence. Assessments were done in 8 separate sessions, with ≤50 images to assess in every session.

To label an ear with a reference diagnosis, a requirement of minimum three panel members choosing that specific diagnosis was set. This procedure rendered a reference diagnosis for 311 ears. Among them, there were 38 cases where a majority of the panel members had chosen the Pass alternative and since that was not possible for the CNN, these cases were discarded at this step. The remaining 273 ears qualified for the CNN analysis will in this thesis be referred to as dataset 1. Across the diagnoses, the number of cases varied, and for acute otitis media there were only two. CNNs require data for training and two images for a class

is not enough. Because of that, decision was made to merge the groups of AOM, OME and CSOM to a class of Pathological. The final classes contained 183 Normal, 46 Wax and 44 Pathological.

For each ear the panel members assessed, they were also asked to rate their confidence in the chosen diagnosis (unsure, confident or very confident) and their judgement of the image quality (poor, acceptable, excellent).

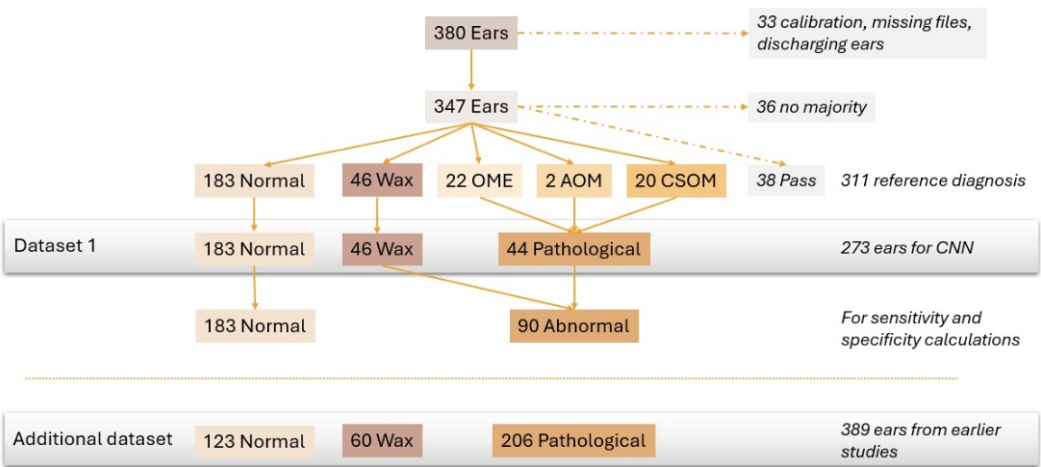


Figure 11. Flow chart describing ears in study III and the distribution of different diagnoses and diagnostic groups. For the 38 Pass images, that was the majority decision of the expert panel. Wax was defined as obstruction of all of the ear canal.
OME = otitis media with effusion, AOM = acute otitis media, CSOM = chronic suppurative otitis media.

For the CNN analysis, the single best image that was previously tagged for each ear was used. The network was trained using 75 % of the images, randomly selected, and tested on the remaining 25 %.

To investigate the premise that CNNs requires large amounts of data to perform well, a separate image bank was introduced. It consisted of 389 tympanic membrane images used in an earlier study by Myburgh at al. [84]. This supplementary data was when used merged with dataset 1 and will in this thesis be referred to as “the combined dataset”.

For image augmentation, we used a method in which the high-resolution original images were used to systematically extract pixels. The extracted pixels from one image were then conflated to a number of low-resolution images (Figure 12). We analysed the performance of the CNN without

any augmentation (original image), as well as for the augmentation factors of 5, 10 and 20. Augmentation factor 5 means that 5 images of low resolution have been created from the original high-resolution image by the procedure described above.

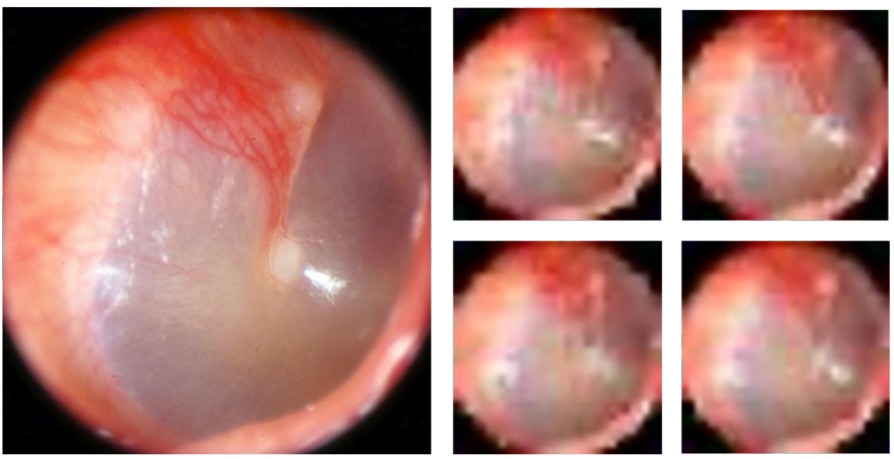


Figure 12. Illustration of augmentation. One high-resolution image is used to extract pixels and create four low-resolution images.

For normalisation, the diagnostic class containing the smallest amount of images decided the level to where the other classes were equalised. For dataset 1, this was the Pathological class containing 44 images. Hence redundant images in the other classes were randomly excluded so that all classes contained 44 images. In the dataset imported from the study by Myburgh et al., Wax was the limiting class [84]. In this dataset the number of images for each class were reduced to 60, to be balanced with this class. The training and testing were done plainly using dataset 1, as well as also utilising the combined dataset for training, but, testing only on dataset 1. The features of normalisation and augmentation was also used, and the combination of these approaches resulted in 16 separate circumstances: datasets (2 different), normalisation (yes/no), augmentation (4 different factors). Lastly, a separate analysis was done on the images where the expert panel had an agreement of Pass.

Procedures study IV

Data collection for study IV was conducted using semi-structured interviews. The interview guide was elaborated according to the steps described by Kalillo et al [85]. The interview questions were discussed within the research group in South Africa and in Sweden. The

interviewer was well-versed with the interview topic from earlier studies of ear and hearing diagnostic methods utilising eHealth and was the leading researcher in construction of the interview guide. The interviewer was also familiar to the environment where the interviews were done from data collections conducted before in similar contexts. To test the interview guide, one pilot interview was performed, audio recorded and transcribed. The pilot interview was used to adjust the interview guide. It was not intended to be included in the data since neither the setting, nor the interviewee were representative for the study aim — it was for practical reasons performed in Sweden. Data was thereafter collected during two separate periods. During the first interview period, five interviews were conducted. The period was used to optimise the interview guide and technique, and queries raised were discussed in the research team. No questions were removed, but one question exploring religiosity in relation to the research topic was added, as well as some follow-up questions, if the initial did not cover a subject in depth. During the second interview period, 12 interviews were held. When a total of 14 interviews were performed, data was considered rich. Male participants and also audiologist were however under-represented, and extra meetings were booked to boost representation from these groups resulting in a total of 17 interviews.

The interviews were conducted in English. The participants were used to communication in English in their daily life, especially in the workplace, but to a varied extent. The interviewer was also used to communicate in English, though not having it as the first language. A research assistant from the City of Tshwane accompanied all interviews. Her role was to facilitate the interview sessions, but she occasionally assisted in clarifying questions. The interviews were sound-recorded, but the quality of the recordings varied a lot due to extensive noise in the surroundings. They were done during the ordinary working day, and participants did not spend any of their free time to participate in the study. During the preparatory phase and interview periods, field notes were made of everything related to data collection and management that was not reflected in the recorded files, and of decisions made along the route.

Demographic data of the participants were collected and recorded in an Excel spreadsheet; year of birth, native language, sex, workplace, profession, length of work life experience, other professional commitments, how often they did take care of ear and hearing symptoms, and also where they turn to gather medical knowledge/ updates. Information was also collected regarding the use of electronic/

computerised equipment in the workplace and of what tools that were available for ear and hearing assessment.

The semi-structured interview covered four research questions.
1. What are the prerequisites for performing ear and hearing diagnostics at the primary healthcare clinics today?
2. How do professionals at the clinics regard their knowledge about AI?
3. What are the attitudes, expectations, qualms and ethical concerns of AI among the professionals?
4. How do the professionals at the public primary healthcare clinics perceive the prerequisite for AI-based ear diagnostics?

The method for analysing the interview data was qualitative content analysis according to Graneheim and Lundman [86-88]. The transcripts of the interviews were considered as the raw material. The first five interviews were transcribed verbatim by the interviewer/ thesis author. The remaining interviews were coarsely transcribed by a collaborator in Pretoria. Because of a high amount of misinterpretation and sections being left out, a second transcription was necessary. It was done by the thesis author, with the coarse version as a template. Data management described in the next sections, was also done by the thesis author.

After transcription, the texts were condensed to a version containing only what was relevant to the actual interview. The raw material contained a section where demographic data was collected, as well as disturbances from ringing phones or staff opening the door or just small talk when having snacks, and these parts were removed. At the end of the interviews, participants were asked if they wanted to read the transcripts from their interview, with an opportunity to clarify or comment. For those ten interested in doing so, these condensed interviews were sent out to them.

In the next step, the condensed interviews were further condensed to a version consisting of condensed meaning units. Dead ends, fillers and iterations of what was just said were removed, if they were not considered to contribute to the understanding of a statement. Also, the lines of the interviewer were removed in this stage, if they were not indispensable for the understanding of the interviewees' answers.

When managing and analysing data, a logbook was kept both for general thoughts raised of the respective interviews when listening to and managing them, and also for records of decisions taken. Sections where it was difficult to hear what the interviewee was saying were marked

together with a time indicator. Some of them could subsequently be solved when listening together with co-researchers and others by reaching out to selected individuals, when the missing parts were technical terms or a specific language. The remaining were left with question marks in the text, but those were only a handful examples.

Before decontextualisation began, the condensed texts were read and short notes created about the general conception from the interview and its content. This was another means of keeping the sense of the whole of the respective interview, in conjunction to the logbook. The condensed meaning units were thereafter imported into the MAXQDA software (version 24.8.0), which was used to facilitate and organise coding. The interviews were coded manually in the software, and no automatic or AI-supported features were used. Two interviews were first coded by the thesis author, using a code close to the text. For triangulation, one co-researcher independently coded the same interviews, and the codes were discussed in the research team to ensure consistency. Thereafter another five were coded by the thesis author using the discussion from the code meetings as a framework. In this thesis, the analysis of these seven interviews will be presented, and the remaining ten interviews will be analysed at a later stage. The codes from the seven interviews were arranged under the four research questions.

Outcomes study I-IV

In the audiometric studies (study I and II), the main outcome was the difference in hearing threshold between the two different audiometers compared. The reliability was also calculated and the time consumption for smartphone audiometry. In study III, the outcome was the accuracy of a machine learning algorithm for middle ear diagnostics, using the diagnoses from an expert panel as the ground truth. The accuracy was also analysed when the different modifications of data was applied. For study IV, since the full material is not yet analysed, the outcome will be presented as a summary of codes related to the respective research questions.

Statistical analysis

In the audiometric studies (study I and II), the paired t-test was used to analyse the difference between conventional audiometry and smartphone audiometry. Significance level was set to $p<0.05$. In study II, significance was also analysed with the Wilcoxon signed rank test.

The paired t-test was used also to calculate reliability for the smartphone hearing test. Calculation of threshold correspondence within e.g. ± 5 or 10 dB (cumulative distribution of signless differences) is an additional analysis for audiological data [89]. With this approach, outliers do not have the same impact on the results, and the presentation offers an overview of the findings.

The performance of the CNN was evaluated by calculating the overall accuracy for each of the 16 combinations of datasets, normalisation and augmentation. The overall accuracy was calculated as the proportion of correctly classified images from the total amount of test images. The accuracy for the respective diagnostic class was also analysed and presented as the true positive rate (TPR) — the proportion of correct classification for each class separately. Sensitivity (also called recall) and specificity were calculated by dividing the images into normal or abnormal (Pathological and Wax).

Ethical considerations

All studies in this thesis are approved by the ethics committees at the University of Pretoria and at Umeå University except for study I in which all parts were performed in South Africa, and it only has the approval from Pretoria. The approval numbers in Pretoria are for study I, II and III 102/2011 and for study IV 587/2023. In Umeå, the approval numbers are for study II and III 2016/257-31 and for study IV 2024-00163-01.

Participation in the studies was voluntary, and participants were enrolled after signing an informed consent form. Study III enrolled participant ≥ 1 year and the parents or legal guardian signed the consent form for under-aged participants. In study I-III, participation in the studies was anonymous and no personal data was collected. The handling of data with potential disclosure of identities in the interview study (study IV) is described below. There was no economical compensation for participation in any of the studies, but participants in study IV were offered a light snack during the interview.

The risk of harm in conjunction to the data collection was considered low. All individuals contributing to the data collection (audiologists, audiology students, thesis author) were well acquainted with the equipment and test procedures, having knowledge of moments with potential risks, including inserting earbuds and ear specula into the ear canal, and testing with loud sounds.

Performing diagnostic tests comes with the possibility of finding pathology. All hearing testing (study I and II) was conducted in collaboration with the University of Pretoria as part of the clinical practice for students in audiology. For participants identified with hearing loss, there was a structure already established for further examinations and treatment. In Study III, images of tympanic membranes were collected. During the collection of these images, in some instances only the thesis author assessed the images, and at instances also an audiologist or audiology students. None of us were experts in ear diagnostics, but as soon as any abnormality was noticed, the person was told so and advised to have a check-up. When possible, they were also assisted in doing so immediately at the clinics. Healthcare at the governmentally run clinics in South Africa is for free for people not having a private healthcare insurance. Advice to have a check-up is therefore not entailed with an economic cost but still requires some of the person's time.

In study IV, confidentiality was considered the highest priority. At all clinics, interviews took place in rooms with closed doors to ensure this. Because of the few numbers of participants in this study, the amount of presented demographic information is kept low. Would we disclose the clinic, occupation, age and clinical experience on a detailed level, there would be a risk of revealing identities. Furthermore, the managers of the clinics know what staff members that participated. Instead, the background information needed to understand the context is presented in general terms and most often on the group level. When data was transferred to the MAXQDA software all geographical cues were masked even when it was just references to e.g. their referral hospitals.

The audiometric application investigated in study I and II (hearTest) was prior to study I created at the University of Pretoria. Before the data collection for study II commenced, the hearX group company was founded and the application became owned and further developed by the company. The application is today a commercial product. There are co-authors in the publications that were linked to the company. De Wet Swanepoel is scientific advisor and one of the founders of hearX with equity interests. Herman Myburgh is also a co-founder of the company and scientific advisor. Claude Laurent was a scientific advisor and consultant to the company at the time of the study. The author of this thesis was never involved in the company, and data collection was done by the thesis author independently without involvement from the company. After the foundation of the company, the persons involved have not been involved in data analysis.

Results

Participants

Included in the studies were in total 303 participants, whereof 63 participated in both study II and III (Figure 13). The age of the participants spanned from 3 to 88 years. Of the included participants, 35 of them were younger than 18 years and they participated in study III. The majority of the participants were female, across the studies 65 %. Males constituted 30 % of the participants and in 5 % data of gender was not collected. Recruitment was done at 12 different clinics or health promotion events.

In study I, two participants were excluded because of low concentration, bringing their babies into the examination room (Figure 13). Two separate ears were also excluded due to unilateral hearing loss >40 dB. From study II, five participants were excluded because of inconsistency in thresholds, misunderstanding of the procedure or problems with fitting of the headphones. Excluded from CNN analysis in study III were images of 20 ears used for calibration of the expert panel, 7 ears because of technical reasons and handling errors and 6 discharging ears. From study IV, all data was included and none of the participants wanted to withdraw during the interviews.

Hearing tests

Threshold comparisons between methods

Enrolled for the audiological studies were a total of 159 participants. Study I had 64 participants in the soundproof environment and 30 in the noisy environment after exclusion of 2 participants (Figure 13). Study II had 63 participants in the only environment, whereof 5 were excluded, keeping 58 for analysis. A participant could for each hearing test contribute to a maximum of 10 hearing threshold (2 ears, 5 frequencies) discounting the repeated measurement at 1.0 kHz for reliability calculation. Measurements missing a corresponding threshold with the other methods could not be compared. In study I, 924 thresholds were available for analysis and in study II, 467 comparisons could be made.

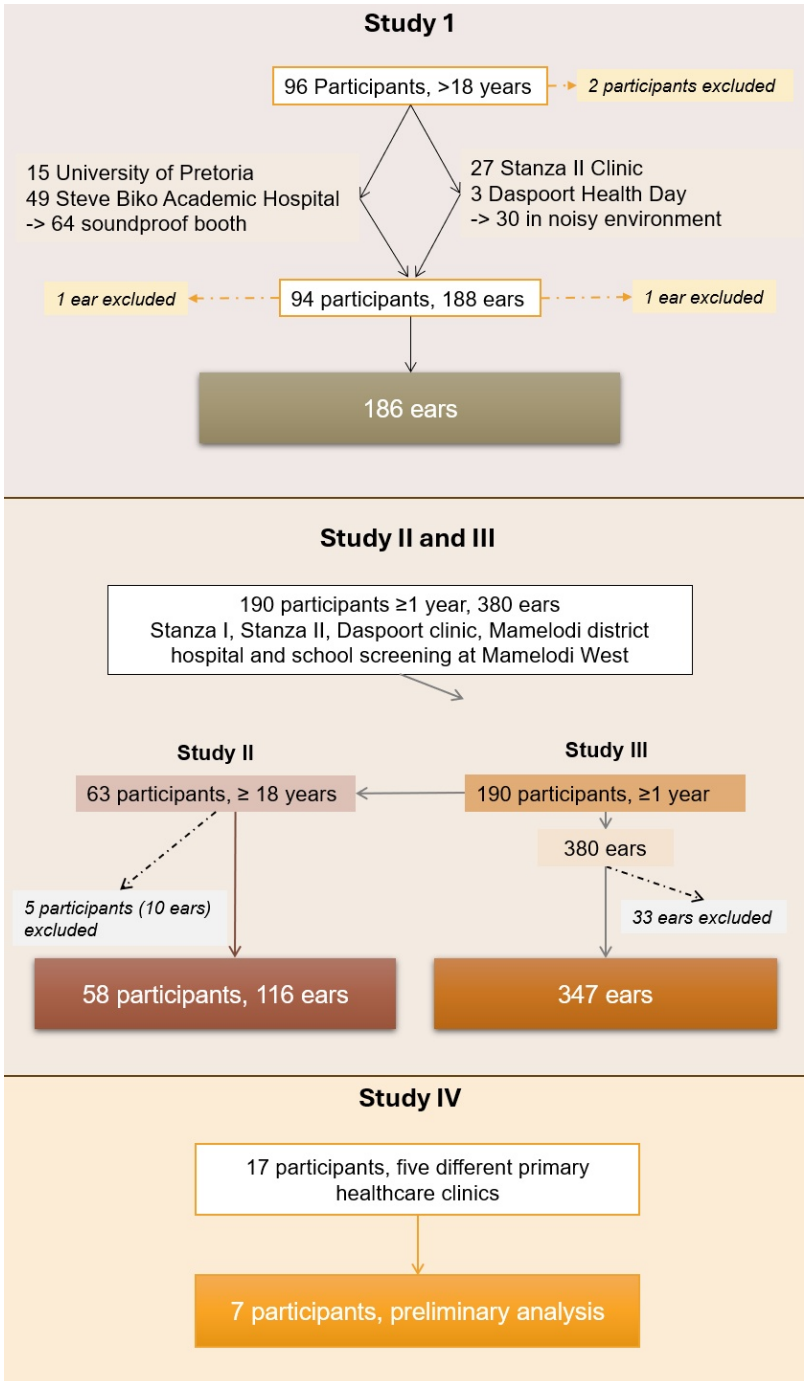


Figure 13. Included and excluded participants and ears for study I-IV and sites for recruitment.

Operator-controlled smartphone test

In study I, in which the smartphone hearing test was operator-controlled, only analysis of thresholds >15 dB with conventional audiometry were further analysed[1]. From this follows that for the soundproof environment 29.5 % (n=186) of the thresholds were included for further analysis and in the noisy environment 67.2 % (n=198) of the thresholds. The results below are for these thresholds.

In the soundproof booth, the mean difference between conventional and smartphone audiometry (smartphone thresholds subtracted from conventional) was 1.6 dB (SD 9.9), implying a need for higher stimulus with conventional method as compared with the smartphone test (Table 3). In the noisy environment, the mean difference was -1.0 dB (SD 7.1) meaning the opposite, a need for higher stimulus with smartphone. The correspondence within ±10 dB was in the soundproof booth 80.6 %, spanning between 72.7 % and 87.2 % across frequencies. In the noisy environment, it was 92.9 % with a span between 88.6 % and 97.7 % across frequencies.

In the soundproof environment, when frequencies were analysed separately, there were statistically significant differences ($p<0.05$) between methods for all of them, except at 4.0 kHz (Table 3). The differences varied between -3.1 dB (at 8.0 kHz) and 5.2 dB (at 0.5 Hz). For all frequencies except at 8.0 kHz, the hearing thresholds were higher with conventional method compared with smartphone. In the noisy environment, the only statistically significant difference between methods when analysing frequency-wise was at 4.0 kHz, amounting to -3.6 dB ($p<0.05$, SD 6.3).

In the published article, only thresholds >15 dB according to conventional method (smartphone thresholds could be at any level) was included. Additional analysis was made including thresholds >15 dB with both methods. In the soundproof environment the mean difference for these thresholds was 0.7 dB (9.7 SD, n=170) and in the noisy environment -1.2 dB (SD 6.9, n=193).

[1] In the published study I, some errors regarding the numbers in the tables and in the body text are unfortunately incorporated. In this thesis, the corrected numbers were used.

	Study I Operator-controlled test Soundproof booth	Study I Operator-controlled test Noisy environment	Study II Self-test Noisy environment
Mean difference (dB)	1.6	-1.0	-2.2
SD	9.9	7.1	10.0
n participants	64	30	58
n threshold	186[a]	198[a]	467
0.5 kHz	5.2 (SD 9.0)*	–1.6 (SD 6.4)	–3.1 (SD 10.9)*
1.0 kHz	3.9 (SD 8.8)*	–0.4 (SD 7.5)	–1.2 (SD 8.8)
2.0 kHz	4.2 (SD 8.7)*	1.2 (SD 6.6)	0.9 (SD 8.6)
4.0 kHz	1.0 (SD 10.1)	–3.6 (SD6.3)*	–3.3 (SD 10.8)*
8.0 kHz	–3.1 (SD 9.9)*	–1.0 (SD 7.7)	–5.4 (SD 9.8)*
Thresholds within ±10 dB (%)	80.6	92.9	80.1

Table 3. Differences in hearing thresholds between methods (smartphone thresholds subtracted from conventional) overall and frequency-wise and correspondence within ±10 dB.
**=p<0.05*
a=thresholds >15 dB with conventional method.

Smartphone self-test

Updating the smartphone hearing test to a self-test and evaluating it in the noisy primary healthcare environment in study II generated a mean difference across frequencies of -2.2 dB (SD 10.0, $p<0.05$), implying a requirement of higher stimulus with the smartphone as compared with conventional audiometry (Table 3). Statistically significant differences were found at 0.5, 4.0 and 8.0 kHz varying between -5.4 and -3.1 dB. No statistically significant differences were found at 1.0 and 2.0 kHz. The absolute difference was across all frequencies 7.7 dB (SD 6.7). The correspondence within ±10 dB was 80.1 %, spanning between 69.4 % and 88.8 % across frequencies. The sensitivity for the smartphone to detect a hearing loss of >40 dB in one ear was 90.6 % (95% CI 75.0-98.0) and the specificity 94.2 % (95% CI 84.1-98.8).

Reliability and time requirement

Analysis of reliability for the smartphone hearing test calculated by comparing two separate measurements at 1.0 kHz in the same ear showed small but not statistically significant differences ($p>0.05$) for the operator-controlled test in both environments as well as for the self-test. The time required for performing a hearing test with smartphone was for the operator-controlled test in the soundproof booth 296 seconds (SD

59) and in the noisy environment 355 seconds (SD 48). For the self-test, time consumption was larger, 512 seconds (SD 203).

Machine learning for middle ear diagnostics

Expert panel

The expert panel examined 347 cases with tympanic membrane images. The requirement for ascribing an ear with a reference diagnosis was a minimum of three panel members agreeing on the diagnosis. That resulted in 36 ears being disqualified by this criterion (Figure 11). Thirty-eight images were assigned the majority decision of Pass. For the remaining 273 ears, the expert panel had fully agreement of the diagnosis in 49 % of them. In 24 %, 4 of the panellists had chosen the same diagnosis and in 27 %, 3 of the 5 panel members agreed. For the respective diagnosis, 5 out of 5 panellists agreed in 49 % of Normal, 59 % of Wax, 0 % of AOM, 4 % of OME and 75 % of CSOM. They rated the quality of the images as acceptable or excellent in 67 % of the cases and stated they were confident or very confident in the diagnosis in 80 % of the cases. The 273 images were distributed in the diagnostic classes as 183 Normal, 46 Wax and 44 Pathological.

Diagnostic performance of the CNN

The diagnostic performance of the CNN was evaluated for overall accuracy and for the three diagnostic classes separately (Normal, Wax and Pathological). The same was analysed when applying normalisation and augmentation and when adding more data (combined dataset). For all analyses that included the combined dataset, it was used to train the CNN, but testing was always solely made on dataset 1.

Overall accuracy

For dataset 1, the overall diagnostic accuracy was 1.00 when no modifications were used (Figure 14). The accuracy dropped when normalising and/or augmenting data, and/or adding more data. Dataset 1 had the highest mean accuracy across all augmentation factors, 0.98 and the lowest was seen when dataset 1 was normalised, 0.91.

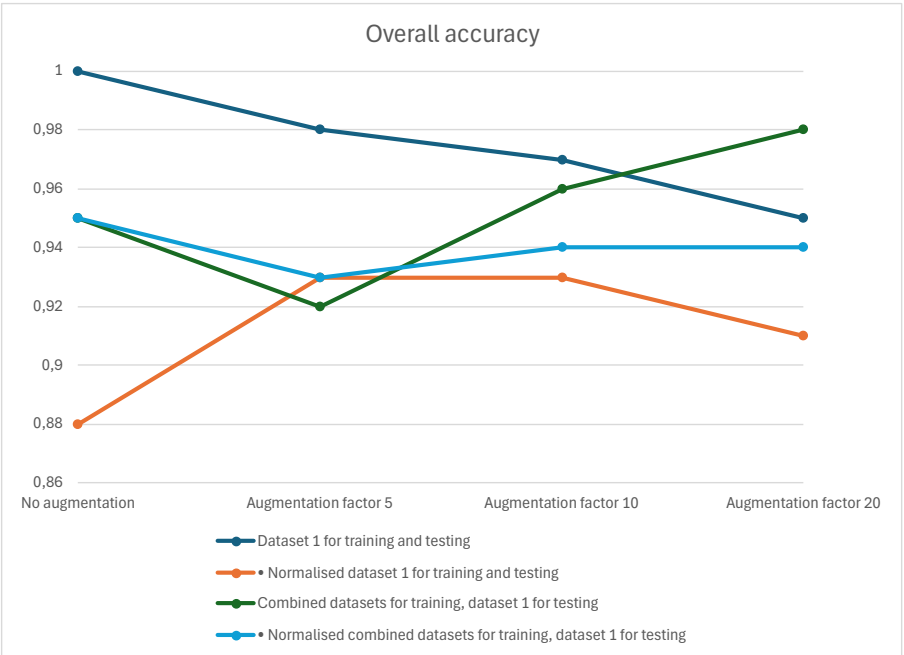


Figure 14. Overall accuracy for each dataset at different augmentation factors. In the published study III, figure 5, the bars showing this are incorrectly arranged. This figure displays the corrected numbers.

Augmentation had no apparent positive effect on the accuracy across all training and testing combinations (Figure 14). At each augmentation factor, there were training/ testing combinations that reached their peak in accuracy at that specific factor. For dataset 1, not normalised, the accuracy decreased with each increment of augmentation.

Normalisation as a comparison between the original datasets and their normalised counterparts showed for dataset 1 a lower accuracy for all augmentation factors (Figure 14). In the combined dataset the effect was not as apparent, but higher accuracy for the not normalised dataset was seen at augmentation factor 10 and 20 but lower at factor 5. In total, there were no positive effect from normalisation on overall accuracy.

More training data was acquired by utilising the combined dataset. When not normalised, the combined dataset performed better than dataset 1 for the highest augmentation factor, but lower on the rest (Figure 14). For the normalised datasets, there was a positive effect from more data. The accuracy was equalled to that of dataset 1 at

augmentation factor 5, and on the remaining augmentation factors, the overall accuracy was better on the combined dataset.

Performance on separate diagnostic classes

When analysing the performance for the diagnostic classes separately (Normal, Wax and Pathological), the overall accuracy of 1.0 for dataset 1 without augmentation was reflected by TPRs of 1.0 for all diagnostic classes (Figure 15). When the modifications of data were used, the lowest TPR seen was for Pathological, 0.59, in the combined dataset without augmentation. The Normal category was not affected by any of the modifications —normalisation, augmentation or more data.

When the effect of augmentation was analysed on the three diagnostic classes separately, the instances deviated the most, and hence affected the overall accuracy the most, were visualised. Pathological and Wax, both in the combined not normalised dataset, seemed to be such cases, at augmentation factor 0 respective 5 (Figure 15). Augmentation seemed to align the TPRs, with less spread on the highest augmentation factor.

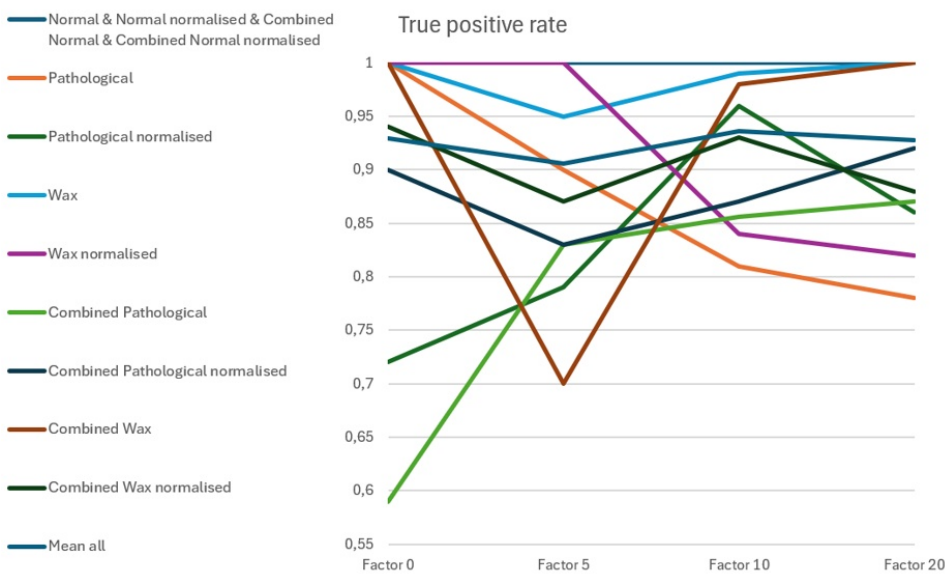


Figure 15. True positive rates for each augmentation factor for the three diagnostic classes (Normal, Wax, Pathological) on dataset 1, the combined dataset and for the normalised datasets. All curves for Normal are merged, since the TPR is 1.0 at all instances.

A consistent gain from normalisation was only seen for the Pathological class on the combined dataset with no instances of decreased TPR (Figure 15). For all other diagnostic classes, there were instances of improvement and deterioration across the augmentation factors for both datasets when normalising.

Adding more training data had a consistent effect only on Wax in the not normalised dataset, with no instances of improved TPR across augmentation factors and deterioration in two of them (Figure 15). For the Normal category, as mentioned in the beginning of this section, there were no effect from any modification, and for the remaining diagnostic classes there were instances of improvement and deterioration across augmentation factors for both the not normalised and normalised datasets.

Sensitivity to detect abnormal ears (Wax or Pathological) was between 93 and 100 % for the different combinations of datasets, augmentation and normalisation. The lowest sensitivity was seen with maximum augmentation (factor 20) for dataset 1. Specificity was 100 % in all instances.

Of the discarded 38 ears with the majority decision of Pass (Figure 11), there were 9 images when all panel members (5 out of 5) had chosen this alternative. These images were transferred to the CNN, to investigate how they would be classified. The setting that previously generated the highest overall accuracy (dataset 1, no normalisation, no augmentation) was used. None of the images were classified as Normal.

Interviews

For the in-depth interviews, 17 participants were recruited from five different community health centres. The interviews lasted between 34 minutes and 1 hour 16 minutes. The participants were nurses, doctors and audiologists. For this theses, seven interviews were selected for analysis. This group included interviews with five females and two males, and all occupations and clinics were represented. The youngest participant was 24 years and the oldest 66 years. The seven interviews were chosen to represent all factors mentioned above as well as possible. The coding procedure of the 7 interviews rendered 827 individual codes. They were summarised under the specific research questions. When the term “AI otoscopy” is used, it refers to imaging of the tympanic membrane with a digital video-otoscope and automated interpretation of

the image by means of AI. This concept was introduced to participants during the interviews.

What are the prerequisites for performing ear and hearing diagnostics at the primary healthcare clinics today?

There were codes disclosing confidence in the basic in otoscopy. Basic procedures, like ear syringing and foreign body removal was considered more difficult. This was due to insufficient training possibilities and insufficient equipment and drugs to perform the procedures. Good spirit within the clinics was often seen, staff helping each other in difficult ear cases. Codes related to hindrances for ear and hearing health touched on facilities in general and resources, like finding a functioning otoscope when needing it or having the proper medication for a certain condition. The limitations on the general level worth highlighting were unreliable internet connection and major difficulties with record keeping and filing. Furthermore, the quality of triaging and sorting of patients was sometimes low, which resulted in long days at the clinics for patients who were sent between different staff members until the person with the right expertise for the specific complaint could see them.

Participant 2: "I think we do have first-line treatment for otitis media and maybe second. But thereafter, with recurrent ear infections, then the guidelines tell you to refer the patient. But we see so many patients with chronic or recurrent type, are you going to send every single one of them? Or a patient that needs syringing. We don't have the resources to do that. But it also just feels very like..., I still send the patient all the way, transport money [for patients], for that. Whereas maybe if we actually have those services here, or better training here."

How do professionals at the clinics regard their knowledge about AI?

Codes spanned from very uncertain in the concept of AI to a high confidence in own knowledge of the technique. A wish for structured introduction or education of AI, if only in the most condensed format could be found. That would weight up for the codes revealing that knowledge of AI was very haphazardly gained and unstructured. In discussions of what AI is, and what can actually be done by AI, misconceptions of its fundamentals and properties were coded.

A wish for some kind of involvement for healthcare staff in the development and implementation of AI was represented in the codes,

both at an early stage and at a hands-on stage when inventions are ready for clinical use. Proper introduction of anything that is new, and proper training were seen as essential for the tool to be used. This was not always a wish at the personal level, but for the professionals in general. Even though there were codes pointing in a reluctant direction when it comes to education, there were also several indicating that there is no way back when it comes to AI, and that healthcare (and the staff) must adapt.

Participant 2: *"Even if it's like a short course or something, just expose us so we know what's this. And I guess, with the way the world is moving, so we're not left behind. Cause imagine, if you don't know anything, in 10 years' time, when AI is more advanced and there's AI in most facilities, you won't notice anything. […] Like I said, I found out about AI through a friend. I was looking for a caption for Instagram."*

What are the attitudes, expectations, qualms and ethical concerns of AI among the professionals?

Overall, the attitude seen was positive and welcoming. There were several tasks specified where AI would be suitable, many of them routine tasks. There were also interest in piloting new solutions although not unison, and interest in contributing to research about AI. From a patient perspective, confidentiality and considerations of how well patients will accept the technique were coded. Many saw possibilities for more accurate diagnosis and better outcome, even though there would be risk of misdiagnosis. From the professionals' perspective, aspects that were subjects of vigilance included the risk of overreliance in the technique and deskilling. A belief in the inherent value in human interactions was a factor with various codes, as something that will limit the extent to which AI will be employed. The concerns towards AI were on a higher level much related to labour market, misuse of collected health data and the accountability question when mistakes occur.

Participant 13: *"It is good [if we could have support in diagnostics], but the problem in South Africa, there is lot of unemployment. It will make things worse. Yona, it's a good thing, but…"*

How do the professionals at the public primary healthcare clinics perceive the prerequisite for AI-based ear diagnostics?

The codes revealed an urge for digital solutions to solve some of the main hindrances to provide good healthcare, not necessarily AI-based

solutions. Furthermore, codes covered problems at the clinic that AI cannot solve e.g. lack of staff, lack of electricity and problems with internet. Would an AI tool rely totally on the abovementioned infrastructure, participants considered it doubtful that an AI otoscope would be useful. Another prerequisite was to ensure its maintenance and provide a proper introduction. When reflecting about education, it was evident that training with the specific tool was seen as indispensable, but regarding whether AI on a more general level, discussing the grounds of the technique, should be taught in healthcare, codes were more disparate.

Participant 7: *"Because there's a lot of times, we've been given something like I think we were given a blood gas machine once for free. They gave us cartridges and everything and then we used it for one month and the cartridges is finished, and they never deliver more cartridges. Now it's in the corner, you see."*

An openness towards AI in general was reflected in various codes and also for ear and hearing diagnostics specifically. Not everyone thought they would personally be the one making the most benefit of AI otoscopy, but still a value of it somewhere in the continuum of care could be seen. Triaging and sorting patients were frequently recurring in codes as such areas where AI could be useful. AI for second opinion instead of referrals was something mentioned as a potential big gain for both patients and personnel. There were however also participants with no full trust in AI for middle ear diagnostics. When participants appreciated their knowledge in regard to using AI otoscopy, there was a variation, not everyone considered themselves knowledgeable enough.

Discussion

Main findings

The technologies for ear and hearing assessments investigated in this thesis are part of the eHealth concept aiming for improved accessibility to ear and hearing healthcare in communities suffering from shortage of basic health services. The interventions were in an early stage of development when the studies were undertaken, and especially study III was of explorative character. The fourth study contributes to chart the applicability of the proposed methods.

When hearing thresholds from smartphone audiometry were compared to conventional audiometry, the mean difference between methods was small for the operator-controlled test (study I) as well as the self-test (study II), between -2.2 dB and 1.6 dB. The machine learning algorithm investigated in study III showed high performance in diagnostics of middle ear pathology. Overall, the accuracy was between 88 % and 100 % across different combinations of training datasets, augmentation and normalisation. Study IV gives indication that the diagnostic methods can be useful in the context where they have been studied for the thesis, even though many aspects must be considered before implementation.

Smartphone audiometry

Differences between smartphone and conventional audiometry

The mean differences between smartphone and conventional audiometry were for the operator-controlled test 1.6 dB (SD 9.9) in the soundproof booth and -1.0 dB (SD 7.1) in the noisy environment. The differences are somewhat larger than the mean difference of 0.4 dB (6.1 SD) between methods seen in a meta-analysis comparing manual and automated air conduction audiometry [90]. The test-retest differences for manual and automated audiometry were 1.3 dB (SD 6.1) respective 0.3 dB, (6.9 SD) in that analysis. The self-test (study II), which was only investigated in the noisy environment, had a mean difference of -2.2 dB (SD 10.0) compared to conventional thresholds. It can still be considered as a clinically insignificant difference, but with slightly higher variability [89,91]. There are few other studies of smartphone threshold audiometry in adults outside of soundproof booths [91]. Our results can be compared to

the study by Renda et al. in which a free, downloadable hearing test application for Android phones used in a quiet room was compared to standard audiometry performed in a soundproof booth. They found the absolute differences to be ≤8.8 dB (SD 8.7) across frequencies [92]. In that study, the bundled headphones were used, with no means to calibrate the equipment, which makes results more difficult to interpret. In a meta-analysis of mobile audiometry (including smartphones and tablets), the mean difference to conventional audiometry was 1.36 dB (95 % CI 0.07-2.66), but the meta-analysis also included studies performed in soundproof booths, which can explain the smaller differences between methods than we had [91]. When we analysed the correspondence within ±10 dB in our studies, it was highest for the operator-controlled test in the noisy environment, 93 %, reaching the expected for measurements on a single frequency when using the same audiometer and procedure [89]. In the two other settings, the correspondence within ±10 dB reached a level of approximately 80 %. All equipment was calibrated prior to the studies, and no major differences should be attributed to unreliable output from the testing devices. Small differences are anyhow permitted in the calibration protocol (ISO 8253-1:2010) and can partly have contributed to the overall differences seen.

Hearing testing in noisy environments

When the smartphone hearing test was studied in the noisy environment, it was seen both with the operator-controlled test (study I) and the self-test (study II) that a higher stimulus was needed for the smartphone than for the conventional test. This finding could partly be attributed to the different attenuation properties for the earphones used in the studies. The Sennheiser earphones used for the smartphone tests are circumaural and not specifically made for hearing testing, while the KUDUwave has insert earphones covered with circumaural ear cups. The self-test thresholds in study II did furthermore show lower correspondence to conventional audiometry compared with a study by van Tonder et al., in which the same smartphone self-test was evaluated, but in a soundproof booth with manual audiometry as reference method [93]. The smartphone thresholds in that study had a correspondence to conventional thresholds within ±10 dB of >88 % across frequencies in an adult population. In study II, only one frequency (2,0 kHz) reached such high correspondence. The surrounding noise is a plausible explanation to the differences between the studies, which was substantial at the primary healthcare clinics where study II was conducted. The headphones used in study II (Sennheiser HD 280 pro) are however

designed with a focus to block out noise, though the shielding effect might not have been sufficient for the setting in which they were used in our study. Furthermore, other distractions were frequent during testing, since there were both other staff and patients in the testing room during the tests. This was an unpredictable factor that might have affected both the conventional test and the smartphone test. The participants also differed between the studies. In the study by van Tonder et al.[93], recruitment was done in audiological clinics and participants might have undertaken audiometry before, being familiar with the procedure. The participants in study II were volunteers from the waiting rooms, and they might have handled the tests in a different way. When compared to other studies of threshold audiometry on mobile devices (including not only smartphones but also tablets), those performed in quiet rooms have not reached the level of ambient noise recorded at three spot checks during data collection in study I (67.3 dBA-68.7 dBA) [92,94-96]. Would tests have been conducted in a private room without doors opening and closing frequently, differences between methods might have been smaller. Furthermore, in an updated version of the smartphone hearing test which includes a noise monitoring function, the permitted sound pressure level is less than our checks were during data collection [97].

Operator-controlled procedure

The results in study I can have been affected by the difference in threshold-seeking approach between the methods. The conventional methods are long-established, but the smartphone method had not been used before. To minimise the risk of random errors, a systematic procedure was used throughout all the testing and followed the threshold-passing procedure used in audiometry. The finding of threshold differences with opposite sign (+/-) in the different environments and across frequencies indicates that the slider-procedure for the smartphone test did not appear to have favoured either of the methods in a considerable way. Such an effect should then have been seen in both environments and across all frequencies. Worth mentioning is that the audiometric test in study I was not intended as an operator-controlled test at first, but instead the participant was supposed to adjust the slider. A pilot test however revealed that the software-design was not intuitive for all participants, and instead we used it in the operator-controlled manner. The development to a self-test in study II has benefits. Except that it is less work-intense for healthcare staff, it is more in alignment with recommendation for audiometry, advising against raising hands as indication for hearing and use of regular repeated

sounds in the testing procedure [18]. In the self-test, the tone presentation and quiet interval in between stimulations were of irregular lengths instead of the constant cycle of tone and pause in study I.

Further development of the smartphone audiometer

After the audiometric studies in this thesis were performed, the intervention has been further developed. The smartphone hearing test has been evaluated for high-frequency audiometry which is of importance when monitoring the hearing during treatment with ototoxic drugs. When smartphone thresholds, established using the same Sennheiser headphones as in study I, were compared to conventional threshold in a soundproof booth, correspondence within ±10 dB (floor effect excluded) was 84 % [98]. The smartphone hearing test has also been studied in the hands of health workers in an operator-controlled manner but utilising the automated features in the self-test, with high sensitivity and specificity in detecting hearing loss [99]. Today, the self-test is used as a low-cost audiometer in clinical settings as well as in research protocols [100].

Machine learning for middle ear disease

Overall CNN accuracy

Study III explored the accuracy of a CNN for diagnostics of middle ear disease, and how augmentation, normalisation and more training data affected the diagnostic performance. For all instances tested, the overall accuracy was ≥88 %. This finding is on par with the performance of many other neural networks tested with different approaches. Livingstone et al. had an overall accuracy of 84 % for 3 diagnostic classes when using a CNN on otoscopic images [101]. Myburgh et al. had for a neural network an accuracy of 87 % for 5 diagnostic classes [84]. One approach to reach higher accuracy is to test several networks and thereafter use the best performing model for further analysis. Cha et al., as an example, tried 9 different CNNs, reaching an accuracy of 92 % for the best performing model when discriminating between 6 different ear diagnoses (and slightly higher for an ensemble model) [102]. Khan et al. tested 5 different CNNs, finally and used the one called DenseNet for further analysis and reached an accuracy of 95 % when classifying tympanic membrane images into 3 categories [103].

Modification of data

In study III we used one CNN but tested different combinations of augmentation and normalisation to find the highest performing setting and the effect of additional training data. The effect of normalisation has not been evaluated specifically in middle ear studies before, even though it has been described as a parameter when experimenting to find higher performing models [103]. The utilisation of augmentation is more frequently used, but again as a preparation for CNN testing, and the effect of the modification as such is seldom analysed [101,103-107].

Normalisation in our study did not only equalise the number between classes for more balanced training but also the testing data. Since we had an overweight of Normal images, the normalised session implies a lesser proportion of these images. Provided Normal was easier for the CNN to recognise, indicated by the 1.0-accuracy at all augmentation factors, a drop in overall accuracy would be expected when normalising, because there are fewer of these easy cases. For dataset 1, this was seen for all augmentation factors. The tendency was seen also in the combined dataset, but not as pronounced.

Augmentation of training data had no apparent positive or negative effect on the CNN performance. The different combinations of training data had zig-zag appearance for overall accuracy, except for dataset 1 which performed worse the more the training data was augmented. Its accuracy could not rise since the vantage point was 1.0, but why it showed lower performance for every step of augmentation is difficult to explain and chance might have had an effect. When analysing the effect of augmentation on the separate diagnostic classes, the curve for the mean across classes only had minimal fluctuations.

The effect of added training data (combined dataset) can be summarised as a tendency towards deterioration on the not normalised dataset and a tendency towards improvement with normalisation. It is difficult to explain why the larger amount of training data would have opposite effect on the not normalised and normalised dataset, but the differences are small in absolute performance and might have been affected by chance.

Triaging and screening

From the perspective of triaging and screening, it can be of value that the

classes were imbalanced. In a screening situation it is expected that most ears are normal and hence the sample mirrors a plausible clinical situation. A network trained to sort out normal ears with high accuracy can be handled in a manner that everything else is redirected to the trained staff for further management. In communities where trained personnel are limited, less trained or untrained personnel might do the triaging. Even though machine learning approaches of 2025 are more sophisticated than the one used for study III, the results from the study still give an indication of the usefulness of the technique for sorting patients. Sensitivity for detecting abnormal ears was high, and specificity was 100 %, indicating that if used for triaging, nearly all diseased ears would be sorted for further assessment, and patients can furthermore trust that their ears are healthy when they are sorted as such. Another interesting finding for the same argument is that none of the Pass-images were classified as Normal.

The reasoning above should be seen as an indication, and the numbers are not fully transferrable to a clinical situation. At first, the users of the system need access to the same otoscope as was used in our study. Furthermore, only images actually displaying the tympanic membrane or obstructing wax were included in our analyses. Would someone without training in otoscopy use the system, there is a risk of capturing images of other structures of the ear than the tympanic membrane. Since such images were not included in the study, and also unexpected findings can appear, there are uncertainty inherent in the method. There are examples from the auto-driving car industry where cars were trained to recognise e.g. bicycles and give way, but not bicycles with bags on the handlebars, leading to accidents [108]. The classification of the Pass-images as not Normal is good as a start, but clinical studies are necessary. It is also important to remember that the desired features of a diagnostic tool vary from place to place. The disease panorama might differ, and training data collected in one setting might not be relevant in another.

The reference diagnosis

The prerequisite for supervised learning to label images (assign a reference diagnosis) for CNN training is one of the major obstacles hampering the development of them, because it takes a lot of work [41]. If training data is unreliable, the output will also be so. In study III, we used a panel of five otologists to label the images and a minimum of three of them to agree to ascribe an image with a diagnosis. This procedure rendered one of the most solid reference standards in studies of machine learning for middle ear diagnostics. Often only one or two

doctors' assigned diagnosis constitute the ground for the reference [84,101,104,105,109-111]. Except from the cost in terms of work needed for a robust standard, it also lessens the amount of data available for training. In a study of automated diagnostics of tympanic membranes, the images available decreased by as much as 78 % when using only images with full agreement between three experts and a certain level of confidence in their diagnosis [58]. In our material, there were no majority in 36 of the 347 images (10 %). The decision to have three out of five panel members was because the study was a development of a method and it was considered good enough as a standard, but without discarding more images than necessary. To be able to include at least part of these images, a discussion in the expert panel could have resulted in a consensus decision. Since the group consisted of five members from different parts of the world it was however considered hard to carry out.

For the 273 images transferred to the CNN, the agreement within the expert panel varied across the diagnoses. For OME, panellists had full agreement only in 4 % whereas for CSOM there was full agreement in 75 %. This finding is aligned with what is known about OME diagnostics, that it is difficult [30,32,58]. Our findings show that OME is a challenging diagnosis even for experts. Images were the only source of information in our study and similar findings were seen for neurotologists also assessing tympanic membrane images. With eight diagnoses to choose between, the inter-rater diagnostic agreement, calculated as kappa scores, was between 0.12 and 0.31 across diagnoses, excluding the diagnosis "tube in place" [112]. This corresponds to slight to fair agreement. Assessment of the retina is also based on images, and ophthalmologists like the otolaryngologists in our study do not agree all the time [113]. It seems difficult to expect a reference diagnosis based on complete agreement.

To get more training data, we could have asked the panel to only choose between the three classes that were subsequently used in the CNN (Normal, Wax and Pathological). There would presumably have been more images with majority decision if the panel members only had to agree that there was *some kind of pathology* but not to specify which kind. We did not know until the results from the expert panel work was summarised that some of the diagnoses was that sparsely represented. But even if the representation would have been perfectly even with no Pass-images, it would have been only 69 in each group and still far from sufficient to train a CNN.

The quality of the images in dataset 1 was by the expert panel rated as "poor" in 33 % of the cases. The images were collected by one person solely, well acquainted with otoscopy and the specific high-end video-otoscope used, and still one third of the images were of poor quality. This is worth noticing when envisioning a future of AI diagnostics. The users will probably have otoscopy as a task of many and maybe not the time to refine the otoscopic technique. Would we have included also images with the majority decision of Pass in the calculation, the proportion of poor-quality images would have been even higher, since poor image quality was one of the reasons the panellists reported for choosing the Pass alternative.

Computer vision for diagnostics

Computer vision for diagnostics in healthcare has some general challenges but there are also specific issues related to ear diagnostics. Among the general problems are those related to the pure technical and methodological features, like available and reliable input data as discussed above, and the risk for bias if training data is not representative for the population [41,114,115]. Often discussed and important from many aspects is also the black box problem of machine learning algorithms [43,116]. When no explanation to the algorithms' decisions is offered, it is more difficult to consider the suggestion and to learn something from the case. Among the general issues are furthermore those practical and ethical ones, like how the technique could be used in the best way to support healthcare staff (and not force the staff to adjust according to the technique) and who would be accountable when mistakes happen. The favourable outcome AI has capacity to deliver will not happen by itself [117-119]. We all have to be active to make the best use of the technique. There are disciplines more or less relying on images solely, but for otitis media, medical history and clinical parameters are indispensable information in decision-making for a diagnosis. An AI-diagnosis based on images solely is therefore challenging. It is known that tests for middle ear effusion enhance the diagnostic accuracy of otitis media [120,121]. It is however also known that even when tympanometers and pneumatic otoscopes are available, they are not fully utilised [122-124]. Studies investigating the possibility to use the information about middle ear effusion in machine learning algorithms have been made. Stephens et al. combined images extracted from videos with results from tympanometry in two algorithms. The diagnostic accuracy was higher when the combined information was used, rising from 78 % to 82 % when distinguish normal ears from abnormal [125].

More recent machine learning algorithms for middle ear diagnostics have approached some of the difficulties we encountered in study III. The main problem was the limited data for algorithm training. Among the largest datasets for development of a CNN for middle ear disease had Zeng et al., using 20542 images [102]. With that amount, they were able to have 8 different diagnostic categories and reached an accuracy of 96 %, but the labelling process in that study is not well described. Wu et al. had for their CNN models 10703 images for training and another 1500 for testing [111]. With 3 diagnostic classes (Normal, AOM, OME) they reached an accuracy of 98 % for the best performing model. Zhong et al. had 20122 images and an accuracy >97 % for two different CNNs when distinguish between normal tympanic membranes, AOM and OME [126]. In the later study, it is however uncertain how they managed to create this large image bank, the same ear might have been included several times. If the same ear is represented in many images, the accuracy will be falsely too high [44]. How they created their reference diagnoses is also not described [126]. These large image datasets are, however they are created, in any event exceptions. A systematic review of machine learning approaches for middle ear diagnostics based on tympanic membrane images found that the median amount was 507 images. It is more than we had in study III, but it seems like acquisition of labelled images is a continuous challenge [107].

The expert panel had a series of images from each ear to assess to make diagnosis, but the CNN only the single best one. In the preparatory phase of study III, we experimented with a stitching technique with the idea to utilise the full series from each ear to create one image that covered as much of the tympanic membrane as possible. This, however, turned out to be too complicated at the time for the CNN analysis. Other researchers have used videos of the tympanic membrane to create stitched images, and the authors in one study used a CNN to do the stitching itself [127]. The network identified certain areas of interest in the membrane to stitch and when using these images, otolaryngologists had greater diagnostic capabilities compared with images stitched in a basic way using all information in the images.

In study III, we also tried to find solutions to utilise the smartphone as an otoscope instead of using a costly external video-otoscope. The images collected however turned out to show a too small portion of the tympanic membrane to be useful. Smartphone otoscopy in conjunction with AI has been evaluated after our study was conducted, both utilising the camera of the phone and with external plug-in devices [106,111,128].

The accuracy for most proposed models using CNN or other machine learning approaches is well above what clinicians reach [107,129,130]. Most studies are however designed for model development and performed in experimental settings not comparable to a clinical situation. One difference from a clinical situation is that in our study as well as the aforementioned studies by Wu et al. [111], Zhong et al. [126] and Stephens et al. [125] (in Zeng et al. [102] it is not described what images that were chosen) and most other, only images nicely visualising the tympanic membrane without wax or other disturbances are included. Furthermore, to enable the analyses in our study, a great effort of manual work was necessary. Disregarding the selection of images which was manually done, preprocessing by cropping the images free from black borders was also needed before data could be handled by the CNN. Studies of machine learning for middle ear diagnostics evaluated in a real clinical setting are still missing and are necessary to be able to evaluate the true clinical value.

Enhanced diagnostics in underserved settings

The underlying healthcare system

In study I-III methods for diagnostics of ear and hearing disorders were investigated. Innovations based on new technology are often highlighted as means for better healthcare in low-income countries [40]. Access to diagnosis does however not necessarily improve health and quality of life for patients. Available treatment is needed as well. For ear disease and hearing loss, there is evidence that treatment, whether it is surgery, hearing aids or other rehabilitation measures, enhances health and the quality of life in many aspects [20,131,132]. But the bulk of patients with hearing loss cannot benefit from the rehabilitation services. Globally, only 17 % of people in need of a hearing aid actually have one. In the African region, the number is as low as 10 % and also in high-income countries the majority is without [133]. There are several reasons for this situation, whereof cost is one of them [133,134]. Increasing the number of diagnoses by better access to diagnostics might therefore not necessarily improve the ear and hearing-related health. In an era of great enthusiasm towards AI for solving healthcare provision shortcomings, the support of the underlying healthcare system must not be forgotten [135]. Making diagnosis but not providing further treatment, or patients are unable to follow referrals, is in many aspects just a cost. It is also important to be vigilant to not create displacement effects [35].

Underscoring the importance of keeping a broad perspective is the fact that affordability is a reason to not even in the first place seek healthcare for ear symptoms [136]. Enhanced diagnostic possibilities do not solve that. Furthermore, discharging ears, as a symptom of CSOM, is not always something that bring neither adults nor children to the healthcare centre, even when care is available [137,138]. Alternative healing is another contributing factor for people to not seek the healthcare centres. Reasons for turning to alternative medicine can be lack of funding or ignorance [136]. Misbelief does also exist, and people instead consult locally elders or other medicine men and sometimes trying remedies doing harm [137,139]. Fighting this is not a challenge confined to ear and hearing healthcare, but something that takes broader perspectives with improved education and empowerment of healthcare personnel. Hearing loss can furthermore be a stigma, as a widespread notion of cognitive decline and a sign of aging, making people hesitate to seek help [140,141].

Doing the right thing, results from study IV

Findings in study IV open up for discussion whether reengineering the management process for ear and hearing patients would make a better contribution to the overall ear and hearing health, instead of introducing new diagnostic tools. Decisions of who to take care of the respective symptom within the clinic, update-trainings in how to treat basic conditions and guidelines when to refer patients were subjects of areas to improve seen in the codes, even though there also were codes reflecting a satisfying organisation. Updates on certain procedures was also asked for, to reduce referrals. Referring a patient might, instead of offering specialised care as intended, lead to no follow-up, due to low compliance. In a study from Malawi, only 3 % of parents to children (5 out of 150) referred to hospital for otolaryngological service actually followed that recommendation [142]. Through interviews, the researchers could uncover some explanations to this. At least some of the reasons found in their study should be applicable to patients attending the clinics in study IV and other clinics in unprivileged areas, e.g. indirect costs and limited knowledge of ear and hearing health. Two studies from South Africa have investigated reasons for not attending scheduled visits. At an otolaryngological clinic following patients after cancer treatment, transport problems was the most common cause [143]. At an outpatient hospital clinic, the most common reason given was unawareness or confusion of the date [144]. However, most of the non-attending patients in the latter study (80 %) could not be reached, and there might be several reasons still unknown.

The interviews revealed that the staff is dealing with many issues related not primary to patient care but solving problems of administration or everyday maintenance of the facilities. The results underscore that the support for the basic service is crucial and must not be neglected. What was further seen in many codes was frustration related to lack of repairs of existing equipment e.g. automatic machines for blood pressure measurements and Hb analysis, which made the machines end up unused as soon as a problem occurred. There is support in previous studies that maintenance is something often overlooked and that stakeholders do not calculate on the real cost including this when planning new procurements [145]. Devices with limited hidden costs (including maintenance) and training for the staff would therefore be of great value, if tools should be used and not just being an expense, or "put in the corner" as one interviewee stated. Lack of internet connectivity at the clinics was another concern reported. Devices relying on connectivity does therefore also have a risk to end up in the corner.

Knowledge and outcomes of AI in healthcare

Codes in study IV revealed that there were participants with little or no conception of AI, however there were also some with great knowledge. Of participants with little or no conception, some could still theorise about AI, and several codes reflected a welcoming attitude. The positive attitude seen has been described among healthcare professionals and students before in other contexts [63,66,146-149]. If knowledge of the technique is lacking, there is however a risk for misuse of it, which in worst case can harm patients. Pitfalls can be inherent in the technique, (as discussed earlier) and that is a risk even with a welcoming attitude. The sources for gathering AI knowledge could from our codes be described as passively encountered and the content of the sources unverified, much picked up from social media, TV and ordinary movies. This has previously been highlighted as a problem that can be solved by AI education [63]. In study IV, participants showed more interest in education of the specific tool (AI otoscope), than learning AI in a more generalised perspective. In a study by Castagno and Khalifa, it was concluded that knowledge of AI was low, but it was discussed more as a problem related to delay in AI implementation because of risk for resistance, than as a patient risk [147]. The level to be considered knowledgeable of AI in that study was quite advanced. In our study, we did not ask specifically about different elements of AI. However, based on our findings, it is reasonable to assume that participants in study IV would be considered to lack knowledge according to the criterion used by

Castagno and Khalifa. In a study from Iran, the requirement to have accurate knowledge was held less rigorous when participants were asked to describe AI in one sentence. Still about half of medical and paramedical students were unable to do this in a correct way [150]. What would be sufficient knowledge depends on what the perspective is and from who's perspective. In our study, we found codes where participants did not regard their knowledge of AI at a level where they would feel comfortable using the technique in their professional work, which is a practical perspective. It gives information whether a tool is likely to be used or not, but less information of the outcome of the usage. Moreover, a positive attitude is not a guarantee for an intervention to be used. In a study from South Africa, healthcare workers were equipped with an mHealth application to report adverse effects of treatment of multidrug-resistant tuberculosis [62]. Although the participants reported high comfort using it and high usefulness, the frequency of reports was unsatisfying.

Notes from the field

In the published articles and the results section in this thesis (study I-III), accuracy and reliability are the main outcomes, as it should in diagnostic studies. During data collection for this thesis, I did however make experiences that add other considerations about the usability of the methods, factors not that easy to measure and put in a table. One of them is difficulties because of language and communication. When multiple languages are spoken within a country (there are 12 official languages in South Africa), perfect communication in all patient-healthcare provider situations are impossible. I met staff in different positions at the clinics that were willing to help with translation, but if not acquainted with the methods, something might get lost along the route. Instructions for the procedures and results/ recommendations in the most common languages would be of great value to reach accurate results and to involve patients in their care, although all languages can be difficult to cover. Participants in study IV, and other staff members I met, told me that patients came also from outside the country, and no one at the clinic could make conversation. Another missing link is that for the communities where the studies were conducted, we do not know the compliance to follow-up for ear and hearing problems. It would be reasonable to assume that patients attending the healthcare centres for an ear problem specifically are more prone to adhere to a prescription or referral. But using the techniques for screening, when people do not recognise they have a problem, might lower the follow-up rate. At one of the hospitals where data was collected, several patients that attended the

outpatient clinic in the morning were put on the operation program for the afternoon. This probably made more patients undertake the necessary surgery and also minimising waste of diagnostic resources not leading to treatment and lessening of the burden of disease. Another experience was the difficulties in planning. Load shedding (planned breaches in delivery of the electricity due to limited capacity) or safety restrictions could occasionally make it impossible to visit clinics and do testing. Furthermore, I was told patients do not always come when they were scheduled, but at unpredictable dates. To get the patients in most need of the tests to come to the clinic the day tests can be performed is a challenge.

Methodological considerations

Applicability of hearing tests

The applicability for hearing tests with the smartphone in normal hearing people is uncertain, since those thresholds were excluded in study I. In study II only a minority of participants had pure tone average ≤20 dB. Research must focus on, besides developing new techniques, also learn when and when not a certain technique is suitable to use. It might be that monitor normal hearing is inappropriate in a noisy primary healthcare environment, since noise is more likely to affect the results on lower sound pressure levels. However, utilising the techniques when accuracy can be guaranteed would be of great value, reducing the cost and enhance availability of hearing tests and saving patients travel expenses and time for visiting a hearing clinic.

Floor effect

Any comparison when one method is permitted to measure higher or lower than the other would have inherent bias. In study I, only thresholds >15 dB with conventional audiometry were analysed in detail. There are limitations with this comparison even though it was necessary to do a subdivision in some way, because of the different capacity for sound output of the audiometric methods used. The mean differences between methods calculated in the study was the smartphone thresholds subtracted from the conventional. Hence, a positive value means a need for higher stimulus with the conventional method and a negative value means the opposite, a need for a higher stimulus with the smartphone test. In the groups presented, smartphone thresholds could be ≤15 dB when conventional thresholds could not, introducing a systematic bias.

Smartphone thresholds <15 dB induce a risk that the mean difference between methods is falsely too positive. In the soundproof booth, the difference between methods was 1.6 dB. By the additional analysis comparing thresholds >15 dB with both methods (instead of only conventional methods) a shift in the negative direction to 0.7 dB was seen. In the noisy environment the difference also became more negative in this additional analysis, -1.2 dB instead of -1.0 dB, illustrating this effect. If we would have changed the criterion to ≥15 dB for conventional method (instead of >15) for inclusion in the analyses, we could have lowered the proportion of thresholds that were not analysed (71 % in the soundproof booth and 33 % in the noisy environment) without causing extra floor effect, since both methods could measure lower than this. Lastly, conventional thresholds <15 dB, but above this level with smartphone would not cause a floor effect and could also have been included in the analysis. By the subdivision we made, results are however applicable for people with hearing loss because the subdivision was made according to the hearing loss criterion with the reference method.

In study II, the floor (and roof) effect did not have the same impact of the results since the test protocols were designed to use the same minimum (and maximum) levels for both methods. Thresholds at minimum level (10 dB) do still have one kind of floor effect, since a difference between methods might have been discernible if it would have been possible to measure at lower output. Our mean difference between methods (-2.2 dB) might have been slightly overestimated because of this. Threshold, irrespective of method, that reached the maximum output level were excluded. Excluding minimum and maximum thresholds would from a clinical point of view be reasonable. Thresholds close to minimum indicate normal hearing and it is in many instances not necessary to investigate further. Maximum thresholds indicate a profound hearing loss, and such cases ought to be picked up and would need more delicate management.

Figure 16. The kudu antelope has big ears, and the portable audiometer used in the studies is named after it.

Conventional audiometry

The hearing thresholds collected in the noisy environment in study I and II were not

compared to the gold standard in audiometry, manual testing by an audiologist in a soundproof booth, but instead we used a portable audiometer (KUDUwave). There were several reasons for this. One of them was that the portable audiometer was already used at the clinics, and inclusion for the studies could therefore take advantage of already established schedules for hearing testing. It could also be argued to be a strength to compare to the existing praxis. Moreover, transport of participants to an audiological clinic for conventional testing would have required very different logistics than what was feasible. The KUDUwave audiometer is however validated in many studies and provides reliable thresholds also outside a soundproof booth [39,79-82].

Order effect and handling of data

Hearing tests were performed in a sequence. The participants had a few minutes to let focus drift away before the next test commenced. By this proceeding the risk of temporal changes in hearing are minimised, as compared with doing them another day or week. Build-up of impacted wax or acquiring or spontaneously resolving OME could be such factors. On the other hand, a fatigue component might have interfered and participants suffering from tinnitus might need even longer time to be fit for more testing. In study I, approximately half of participants started with conventional audiometry and half with smartphone audiometry to avoid a systematic bias. In study II close to all participants did for practical reasons start with the conventional test and there might have been an order effect.

Since the results from the audiometry studies were manually recorded in an Excel spreadsheet, there is a risk of error in this process. No data was transferred automatically to the digital format. The transfer was however done with great precautions and the risk for error is small and for systematically bias the risk should be minimal.

CNN performance

Since the image bank was constrained, the outcome from testing different combinations of training data, augmentation and normalisation, might have been affected from chance. The network was trained on different images randomly sampled for each of the scenarios. Since every scenario was only run once, part of the effects seen might be associated to this proceeding. In other words, the amount of exposure to different diagnostic classes did probably vary from scenario to scenario

when training the algorithm. As an example, there were 44 images classified as Pathological in dataset 1. Training the CNN on 75 % of them would provide 33 images for training, but since the 75 % were randomly selected from all 273 images, there were probably instances when the amount of Pathological training images was far less. From that follows that a repeated-round approach might have shown a more reliable relation. Another limitation is the under-representation of AOM. The two images in the dataset can only to a minimal extent have had an effect on results. In a clinical situation, diagnostics of AOM is a desirable target, especially in the light of fighting antibiotic over-prescription. More images of this diagnosis are crucial in an image bank. To attain better training, we could also have used part of the images for validation [151]. This would have lowered the risk for over-fitting, the instance when an algorithm has trained to much on a specific dataset and by that is too adapted to it, at the expense of fitness to new images [44,151,152].

There are more reasons to assume that the accuracy in study III is over-estimated and that it would not be as high in a real clinical situation. As discussed earlier, all images were taken with the same otoscope. The fact that test images were from the same otoscope as training images probably enhances the accuracy, and it is not representable for a situation of launching a network like this for clinical use. It would have been of interest to analyse a scenario when the CNN was trained on dataset 1, taking benefit from the solid reference standard, but test it on an external dataset. That would have been closer to a true clinical situation. It would also have been interesting to run the combined dataset for training as we did, but do the testing also on the combined (instead of only dataset 1). It would have been a fairer evaluation of what the added images contributed with. The performance of the CNN did not consistently rise by incorporating more training data. Except from the additional training data, incorporated in the comparison were also the effects of including images captured with different devices and procedures for reference standard. Different devices can vary in lightening, focus and resolution. A decent training dataset should contain images collected during various circumstances and with various devices.

Trustworthiness in study IV

Credibility. The method for data collection was subject for discussion when the study aim was raised. In-depth interviews were considered suitable since we expected variation in perspectives and knowledge that

would be difficult to capture using e.g. a questionnaire. Nuances and thoughts out of the predefined questions would not be possible to explore with such a method. The author of this thesis conducted the interviews and is not from the communities where data was collected. This can be of value from some perspectives and a hindrance from others. The participants cannot expect the interviewer to have the same pre-understanding as someone native would, which could offer more thoughtful explanations. There might although be more questioning towards someone from outside, leading to a reserved attitude which happened at instances. To have our research assistant, who was from one of the communities where data was collected, as the interviewer was discussed. Insufficient knowledge in the methodology and research topic did however rule out this option.

The clinics chosen for data collection were considered representative for the research aim. The South African co-researchers have good insight in the healthcare system and the communities in the City of Tshwane from earlier work, and organised inclusion of desired clinics. The selection of clinics aimed to cover the differences that might exist between them in terms of e.g. population, specific focus areas and challenges. We managed to get participants from the preferred categories of occupation as well as a great span of age and work experience. Women were easier to recruit, reflecting the general composition of the healthcare workforce, but we tried to counter-balance by specifically invite men at the end of the second interview period.

Several measures were undertaken to keep the sense of the context and of the whole. This is important when recontextualising the findings, so that the reality of the interviewee is not reduced [153]. The interviewer listened to many of the interviews close to the interview situation, and also when transcribing them. Interviews were then several times fully read before the coding commenced. The South African co-researcher is familiar with the context for the interviews and has been doing data collection in all selected clinics before. The two co-researchers in Sweden participated as observers in six of the interviews and visited three of the clinics where data was collected. Furthermore, we used member check and ten participants announced they were interested in reading the transcripts, but none of them had any further comments after reading.

Dependability for the presented result should be valid, and consistency throughout the process is considered high. The work has been done stepwise, e.g. condensation of the interviews was finished before coding started. It will be more challenging to keep consistency for the remaining

ten interviews later on, since codes and the organisation of them will colour the process. A measure to mitigate the impact from this proceeding is the logbook where the analysis process is documented.

Transferability. Data for study IV was collected in low-income communities in the City of Tshwane in South Africa, from staff members with different occupation in primary healthcare. The research was focused on tools for ear diagnostics. Results however covered the situation at the clinics in general and across other areas. Some of the findings might therefore be valid more widely but keeping in mind that circumstances at different workplaces can vary a lot and then the applicability of the findings.

The interviews in study IV were conducted in English. To do the interviews in the native language of the participants would presumably have made it easier for them to express themselves. Ten different main languages were although reported among participants, and to find someone knowledgeable in all these languages and at the same time knowledgeable in the research was not possible. Having more than one interviewer could also have been an option, but that would affect the interviews in other ways, e.g. interviewers would have different pre-knowledge. English is however an official language in South Africa and a language widely used in the population when the first language is not an option.

Clinical implication and future studies

Validated eHealth solutions for middle ear diagnostics and hearing assessments can contribute to provide accessible diagnostics in underserved setting. Automated diagnostics based on AI can assist to triage patients with ear and hearing symptoms, but the algorithms must be adapted to the target population and be incorporated in an appropriate tool. When gold standard audiometry is not available, smartphone hearing tests could be an accurate low-price alternative. The technique can improve accessibility to hearing healthcare also in other contexts than those described in this thesis. Waiting times for hearing services are long also in high-income countries, Sweden is one example of that [154]. Furthermore, the distance to the services can be substantial at many places, and the opportunity to utilise a smartphone to undertake a hearing test without having to travel could open up for earlier diagnostics and tele-medical hearing service.

What transpired about attitudes to the new technique in study IV might be useful also when planning introduction of other equipment or changes of practice. Even though the present study focused on ear and hearing conditions, many of the codes covered examples not related to that topic. To further investigate whether AI otoscopy could be used and benefit primary healthcare clinics in unprivileged areas, workshops at site could make a good contribution. A wish to be involved was seen among our participants, but if lacking an experience of AI, it is difficult to take part in discussions. A hands-on exercise followed by interviews could add valuable knowledge of the potential usage of future tools. Furthermore, such an event could also be used to explore the question stated earlier of what sufficient knowledge of AI is, in relation to practical training on selected learning cases. When there are AI tools, technically and practically available for clinical use, studies must be designed to investigate the true clinical outcome for patients. AI research must be an interdisciplinary venture, including all possible partners.

An optimistic outlook forward

Since ear and hearing health are closely intertwined, a reliable low-cost diagnostic solution for examination of the entire system could contribute to early detection of disorders and thereby lessening the burden of the related conditions. Otoscopy and hearing assessments are used in decision of management for the common conditions of CSOM, OME and AOM and when assessing hearing loss, certain middle ear conditions must be ruled out. A diagnostic tool that is accepted by professionals and patients, containing teaching features as well as guidelines of how to handle the diagnosed conditions could reduce barriers for when health workers can examine patients single-handedly. If the tool could combine findings from AI-interpreted otoscopy with the audiogram and present a rational for a recommendation it could increase accuracy of diagnoses and access to care and promote learning progress.

Acknowledgement

Main supervisor Thorbjörn Lundberg, always positive and keeping up the good spirit and finding a way forward even when there seems to be none.

Co-supervisor Professor Claude Laurent, without you nothing of this would have happened. From where do you get all your energy?

Co-supervisors Göran Umefjord and Jenny Andersson for invaluable advice and effort along the route.

South African co-supervisor Professor De Wet Swanepoel always with a good solution, making research look easy. And, for together with the Swanepoel family, showing the best of your wonderful country.

The students at Speech-Language Pathology and Audiology, University of Pretoria, for the good collaboration during data collection.

Christine Louw van der Westhuizen, Charl, Christian and Charlie for your hospitality and for ten years of great friendship.

Leigh Biagio de Jager and Hennie and little Demi, you are a reason to always go back to Pretoria. And for always finding a new fish to braai.

Stacy Jose, never hesitating, whether it's work or a hike an early Saturday morning.

Anna Engström Laurent for amazing support all the time since the early beginning in your kitchen on Teg.

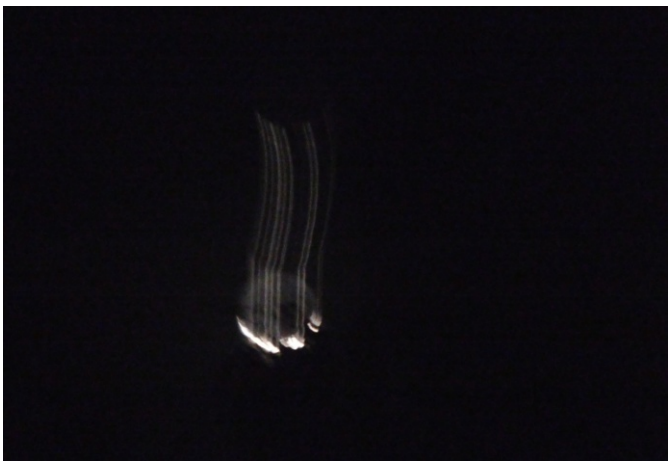
Co- and former PhD students for great company and interesting discussions at the fifth floor.

All smart and sensible people at the Family Medicine Unit.

Maria Sandström for knowing how to solve whatever and for always adding another perspective to the big and small.

To the Hudiksvall people for taking care of the urological patients and for looking after the neighbourhood while I was away collecting data.

The greatest Ingasill and the greatest sisters Johanna and Linnéa.



What would it have looked like if it had looked as though the Earth was rotating?

References

1 Klein, J. O. The burden of otitis media. *Vaccine* 19 Suppl 1, S2-8 (2000). https://doi.org/10.1016/s0264-410x(00)00271-1

2 Collaborators, G. B. D. C. o. D. Global burden of 292 causes of death in 204 countries and territories and 660 subnational locations, 1990-2023: a systematic analysis for the Global Burden of Disease Study 2023. *Lancet* 406, 1811-1872 (2025). https://doi.org/10.1016/S0140-6736(25)01917-8

3 Schilder, A. G. *et al.* Otitis media. *Nat Rev Dis Primers* 2, 16063 (2016). https://doi.org/10.1038/nrdp.2016.63

4 Collaborators, G. B. D. H. L. Hearing loss prevalence and years lived with disability, 1990-2019: findings from the Global Burden of Disease Study 2019. *Lancet* 397, 996-1009 (2021). https://doi.org/10.1016/S0140-6736(21)00516-X

5 Woodcock, K. & Pole, J. D. Educational attainment, labour force status and injury: a comparison of Canadians with and without deafness and hearing loss. *Int J Rehabil Res* 31, 297-304 (2008). https://doi.org/10.1097/MRR.0b013e3282fb7d4d

6 World Health Organization. World report on hearing. (Geneva Licence: CC BY-NC-SA 3.0 IGO, 2021).

7 Livingston, G. *et al.* Dementia prevention, intervention, and care. *Lancet* 390, 2673-2734 (2017). https://doi.org/10.1016/S0140-6736(17)31363-6

8 Lin, F. R. *et al.* Hearing loss and incident dementia. *Arch Neurol* 68, 214-220 (2011). https://doi.org/10.1001/archneurol.2010.362

9 Global Burden of Disease Study, C. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 386, 743-800 (2015). https://doi.org/10.1016/S0140-6736(15)60692-4

10 Pichichero, M. E. & Poole, M. D. Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. *Arch Pediatr Adolesc Med* 155, 1137-1142 (2001). https://doi.org/10.1001/archpedi.155.10.1137

11 Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 399, 629-655 (2022). https://doi.org/10.1016/S0140-6736(21)02724-0

12 Friedman, N. D., Temkin, E. & Carmeli, Y. The negative impact of antibiotic resistance. *Clin Microbiol Infect* 22, 416-422 (2016). https://doi.org/10.1016/j.cmi.2015.12.002

13 Monasta, L. *et al.* Burden of disease caused by otitis media: systematic review and global estimates. *PLoS One* 7, e36226 (2012). https://doi.org/10.1371/journal.pone.0036226

14 Olusanya, B. O. & Newton, V. E. Global burden of childhood hearing impairment and disease control priorities for developing countries. *Lancet* 369, 1314-1317 (2007). https://doi.org/10.1016/S0140-6736(07)60602-3

15 Daniel, E. Noise and hearing loss: a review. *J Sch Health* 77, 225-231 (2007). https://doi.org/10.1111/j.1746-1561.2007.00197.x

16 Nomura, K., Nakao, M. & Morimoto, T. Effect of smoking on hearing loss: quality assessment and meta-analysis. *Prev Med* 40, 138-144 (2005). https://doi.org/10.1016/j.ypmed.2004.05.011

17 Harris, T., Peer, S. & Fagan, J. J. Audiological monitoring for ototoxic tuberculosis, human immunodeficiency virus and cancer therapies in a developing world setting. *J Laryngol Otol* 126, 548-551 (2012). https://doi.org/10.1017/S0022215112000357

18 British Society of Audiology. Recommended procedures for pure-tone audiometry using a manually operated instrument. *Br J Audiol* 15, 213-216 (1981). https://doi.org/10.3109/03005368109081440

19 Cannizzaro, E. *et al.* Exposure to ototoxic agents and hearing loss: A review of current knowledge. *Hearing Balance and Communication* 12, 166-175 (2014). https://doi.org/10.3109/21695717.2014.964939

20 Ferguson, M. A. *et al.* Hearing aids for mild to moderate hearing loss in adults. *Cochrane Database Syst Rev* 9, CD012023 (2017). https://doi.org/10.1002/14651858.CD012023.pub2

21 World Health Organization. Multi-country Assessment of National Capacity to Provide Hearing Care. Report No. 978 92 4 150657 1, (http://www.who.int/pbd/publications/WHOReportHearingCare_Englishweb.pdf, 2013).

22 Mulwafu, W., Ensink, R., Kuper, H. & Fagan, J. Survey of ENT services in sub-Saharan Africa: little progress between 2009 and 2015. *Glob Health Action* 10, 1289736 (2017). https://doi.org/10.1080/16549716.2017.1289736

23 Browning, G. B. in *Open Access Guide to Audiology and Hearing Aids for Otolaryngologists* (ed Claude Laurent De Wet Swanepoel) Ch. Whispered and spoken voice tests and tuning fork tests, (University of Cape Town).

24 Louw, C., Swanepoel, W. & Eikelboom, R. H. Self-Reported Hearing Loss and Pure Tone Audiometry for Screening in Primary Health Care Clinics. *J Prim Care Community Health* 9, 2150132718803156 (2018). https://doi.org/10.1177/2150132718803156
25 Stuart, A., Stenstrom, R., Tompkins, C. & Vandenhoff, S. Test-retest variability in audiometric threshold with supraaural and insert earphones among children and adults. *Audiology* 30, 82-90 (1991). https://doi.org/10.3109/00206099109072873
26 Tashneem, H. in *Open Access Guide to Audiology and Hearing Aids for Otolaryngologists* (ed Laurent Claude Fagan Johan, Swanepoel Dewet) Ch. Otoscopy & pneumatic otoscopy, 9 (University of Cape Town).
27 Laurent, C. in *Open Access Guide to Audiology and Hearing Aids for Otolaryngologists* (ed Swanepoel Dewet Fagan Johan) Ch. Tympanometry, 5 (University of Cape Town, 2024).
28 Pichichero, M. E. & Poole, M. D. Comparison of performance by otolaryngologists, pediatricians, and general practioners on an otoendoscopic diagnostic video examination. *Int J Pediatr Otorhinolaryngol* 69, 361-366 (2005). https://doi.org/10.1016/j.ijporl.2004.10.013
29 Legros, J. M. *et al.* Clinical qualitative evaluation of the diagnosis of acute otitis media in general practice. *Int J Pediatr Otorhinolaryngol* 72, 23-30 (2008). https://doi.org/10.1016/j.ijporl.2007.09.010
30 Buchanan, C. M. & Pothier, D. D. Recognition of paediatric otopathology by General Practitioners. *Int J Pediatr Otorhinolaryngol* 72, 669-673 (2008). https://doi.org/10.1016/j.ijporl.2008.01.030
31 Rogers, D. J., Boseley, M. E., Adams, M. T., Makowski, R. L. & Hohman, M. H. Prospective comparison of handheld pneumatic otoscopy, binocular microscopy, and tympanometry in identifying middle ear effusions in children. *Int J Pediatr Otorhinolaryngol* 74, 1140-1143 (2010). https://doi.org/10.1016/j.ijporl.2010.06.015
32 Lee, D. H. How to improve the accuracy of diagnosing otitis media with effusion in a pediatric population. *Int J Pediatr Otorhinolaryngol* 74, 151-153 (2010). https://doi.org/10.1016/j.ijporl.2009.10.029
33 Socialstyrelsen och Nationella vårdkompetensrådet. Kompetensförsörjning inom primärvården. (2022).
34 Boniol, M. *et al.* The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal'

health coverage? *BMJ Glob Health* 7, e009316 (2022). https://doi.org/10.1136/bmjgh-2022-009316

35 World Health Organization. WHO guideline: recommendations on digital interventions for health system strengthening. Report No. 9789241550505, (World Health Organization, Geneva, 2019).

36 Joshi, R. *et al.* Task shifting for non-communicable disease management in low and middle income countries--a systematic review. *PLoS One* 9, e103754 (2014). https://doi.org/10.1371/journal.pone.0103754

37 Pierre, W. C. Y. & Zee Hui, L. Mobile Audiometry for Use in Ototoxicity Monitoring Programs: A Scoping Review. *Journal of audiology & otology* 29, 168-180 (2025).

38 Munoz, K., Nagaraj, N. K. & Nichols, N. Applied tele-audiology research in clinical practice during the past decade: a scoping review. *Int J Audiol* 60, S4-S12 (2021). https://doi.org/10.1080/14992027.2020.1817994

39 Maclennan-Smith, F., Swanepoel de, W. & Hall, J. W., 3rd. Validity of diagnostic pure-tone audiometry without a sound-treated environment in older adults. *Int J Audiol* 52, 66-73 (2013). https://doi.org/10.3109/14992027.2012.736692

40 Wahl, B., Cossy-Gantner, A., Germann, S. & Schwalbe, N. R. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health* 3, e000798 (2018). https://doi.org/10.1136/bmjgh-2018-000798

41 Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat Med* 25, 24-29 (2019). https://doi.org/10.1038/s41591-018-0316-z

42 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* 521, 436-444 (2015). https://doi.org/10.1038/nature14539

43 Castelvecchi, D. Can we open the black box of AI? *Nature* 538, 20-23 (2016). https://doi.org/10.1038/538020a

44 Varoquaux, G. & Colliot, O. in *Machine Learning for Brain Disorders* Vol. 197 (ed O. Colliot) 601-630 (2023).

45 Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 368, m689 (2020). https://doi.org/10.1136/bmj.m689

46 Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw* 106, 249-259 (2018). https://doi.org/10.1016/j.neunet.2018.07.011

47 Lo, A. H. & McPherson, B. Hearing screening for school children: utility of noise-cancelling headphones. *BMC Ear Nose Throat Disord* 13, 6 (2013). https://doi.org/10.1186/1472-6815-13-6

48 Kam, A. C., Sung, J. K., Lee, T., Wong, T. K. & van Hasselt, A. Clinical evaluation of a computerized self-administered hearing test. *Int J Audiol* 51, 606-610 (2012). https://doi.org/10.3109/14992027.2012.688144

49 Szudek, J. *et al.* Can Uhear me now? Validation of an iPod-based hearing loss screening test. *J Otolaryngol Head Neck Surg* 41 Suppl 1, S78-84 (2012). https://doi.org/10.2310/7070.2011.110089

50 Foulad, A., Bui, P. & Djalilian, H. Automated audiometry using apple iOS-based application technology. *Otolaryngol Head Neck Surg* 149, 700-706 (2013). https://doi.org/10.1177/0194599813501461

51 Handzel, O. *et al.* Smartphone-based hearing test as an aid in the initial evaluation of unilateral sudden sensorineural hearing loss. *Audiol Neurootol* 18, 201-207 (2013). https://doi.org/10.1159/000349913

52 Kam, A. C. *et al.* Automated hearing screening for children: A pilot study in China. *Int J Audiol* 52, 855-860 (2013). https://doi.org/10.3109/14992027.2013.832419

53 Khoza-Shangase, K. & Kassner, L. Automated screening audiometry in the digital age: exploring uhear and its use in a resource-stricken developing country. *Int J Technol Assess Health Care* 29, 42-47 (2013). https://doi.org/10.1017/S0266462312000761

54 Van Tasell, D. J. & Folkeard, P. Reliability and accuracy of a method of adjustment for self-measurement of auditory thresholds. *Otol Neurotol* 34, 9-15 (2013). https://doi.org/10.1097/MAO.0b013e318278c05d

55 Yeung, J. *et al.* The new age of play audiometry: prospective validation testing of an iPad-based play audiometer. *J Otolaryngol Head Neck Surg* 42, 21 (2013). https://doi.org/10.1186/1916-0216-42-21

56 Na, Y. *et al.* Smartphone-based hearing screening in noisy environments. *Sensors (Basel)* 14, 10346-10360 (2014). https://doi.org/10.3390/s140610346

57 Wu, W. *et al.* A new hearing screening system for preschool children. *Int J Pediatr Otorhinolaryngol* 78, 290-295 (2014). https://doi.org/10.1016/j.ijporl.2013.11.026

58 Kuruvilla, A., Shaikh, N., Hoberman, A. & Kovacevic, J. Automated diagnosis of otitis media: vocabulary and grammar. *Int J Biomed Imaging* 2013, 327515 (2013). https://doi.org/10.1155/2013/327515

59 Myburgh, H. C., van Zijl, W. H., Swanepoel, D., Hellstrom, S. & Laurent, C. Otitis Media Diagnosis for Developing Countries

Using Tympanic Membrane Image-Analysis. *EBioMedicine* 5, 156-160 (2016). https://doi.org/10.1016/j.ebiom.2016.02.017

60 Wang, X., Valdez, T. A. & Bi, J. Detecting tympanostomy tubes from otoscopic images via offline and online training. *Comput Biol Med* 61, 107-118 (2015). https://doi.org/10.1016/j.compbiomed.2015.03.025

61 Canares, T. L., Wang, W., Unberath, M. & Clark, J. H. Artificial intelligence to diagnose ear disease using otoscopic image analysis: a review. *J Investig Med* 70, 354-362 (2022). https://doi.org/10.1136/jim-2021-001870

62 Chaiyachati, K. H. *et al.* A pilot study of an mHealth application for healthcare workers: poor uptake despite high reported acceptability at a rural South African community-based MDR-TB treatment program. *PLoS One* 8, e64662 (2013). https://doi.org/10.1371/journal.pone.0064662

63 Lai, M. C., Brian, M. & Mamzer, M. F. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. *J Transl Med* 18, 14 (2020). https://doi.org/10.1186/s12967-019-02204-y

64 Jungmann, F. *et al.* Attitudes Toward Artificial Intelligence Among Radiologists, IT Specialists, and Industry. *Acad Radiol* 28, 834-840 (2021). https://doi.org/10.1016/j.acra.2020.04.011

65 van Hoek, J. *et al.* A survey on the future of radiology among radiologists, medical students and surgeons: Students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over. *Eur J Radiol* 121, 108742 (2019). https://doi.org/10.1016/j.ejrad.2019.108742

66 Scott, I. A., Carter, S. M. & Coiera, E. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health Care Inform* 28, e100450 (2021). https://doi.org/10.1136/bmjhci-2021-100450

67 Blease, C. *et al.* Artificial Intelligence and the Future of Primary Care: Exploratory Qualitative Study of UK General Practitioners' Views. *J Med Internet Res* 21, e12802 (2019). https://doi.org/10.2196/12802

68 Liyanage, H. *et al.* Artificial Intelligence in Primary Health Care: Perceptions, Issues, and Challenges. *Yearb Med Inform* 28, 41-46 (2019). https://doi.org/10.1055/s-0039-1677901

69 Katonai, G., Arvai, N. & Mesko, B. AI and Primary Care: Scoping Review. *J Med Internet Res* 27, e65950 (2025). https://doi.org/10.2196/65950

70 World Bank. World Bank country classifications by income level for 2024-2025.

(https://blogs.worldbank.org/en/opendata/world-bank-country-classifications-by-income-level-for-2024-2025, 2025).
71 Chatterjee, A., Czajka, L. & Gethin, A. Wealth Inequality in South Africa, 1993–2017. *The World Bank Economic Review* 36, 19-36 (2022). https://doi.org/10.1093/wber/lhab012
72 World Bank Group. The World Bank in South Africa. (https://www.worldbank.org/en/country/southafrica/overview, 2025).
73 Gordon, T., Booysen, F. & Mbonigaba, J. Socio-economic inequalities in the multiple dimensions of access to healthcare: the case of South Africa. *BMC Public Health* 20, 289 (2020). https://doi.org/10.1186/s12889-020-8368-7
74 World Health Organization. Country Cooperation Strategy 2023-2027, South Africa. Report No. ISBN: 9789290314257, (Brazzaville: WHO African Region, 2025).
75 Babatunde, A. O., Akin-Ajani, O. D., Abdullateef, R. O., Togunwa, T. O. & Isah, H. O. Review of antiretroviral therapy coverage in 10 highest burden HIV countries in Africa: 2015-2020. *J Med Virol* 95, e28320 (2023). https://doi.org/10.1002/jmv.28320
76 World Health Organization. Global tuberculosis report 2024. Geneva (2024).
77 Statistics South Africa. Statistics South Africa on Quarterly Labour Force Survey. (South African Government, https://www.gov.za/news/media-statements/statistics-south-africa-quarterly-labour-force-survey-qlfs-%E2%80%93-q1-2025-13-may, 2025).
78 Swanepoel de, W., Myburgh, H. C., Howe, D. M., Mahomed, F. & Eikelboom, R. H. Smartphone hearing screening with integrated quality control and data management. *Int J Audiol* 53, 841-849 (2014). https://doi.org/10.3109/14992027.2014.920965
79 Swanepoel de, W., Mngemane, S., Molemong, S., Mkwanazi, H. & Tutshini, S. Hearing assessment-reliability, accuracy, and efficiency of automated audiometry. *Telemed J E Health* 16, 557-563 (2010). https://doi.org/10.1089/tmj.2009.0143
80 Swanepoel de, W., Matthysen, C., Eikelboom, R. H., Clark, J. L. & Hall, J. W., 3rd. Pure-tone audiometry outside a sound booth using earphone attentuation, integrated noise monitoring, and automation. *Int J Audiol* 54, 777-785 (2015). https://doi.org/10.3109/14992027.2015.1072647
81 Storey, K. K., Munoz, K., Nelson, L., Larsen, J. & White, K. Ambient noise impact on accuracy of automated hearing assessment. *Int J Audiol* 53, 730-736 (2014). https://doi.org/10.3109/14992027.2014.920110

82 Visagie, A., Swanepoel de, W. & Eikelboom, R. H. Accuracy of Remote Hearing Assessment in a Rural Community. *Telemed J E Health* 21, 930-937 (2015). https://doi.org/10.1089/tmj.2014.0243
83 Szegedy, C. *et al.* in *IEEE conference on computer vision and pattern recognition* 1-9.
84 Myburgh, H. C., Jose, S., Swanepoel, D. W. & Laurent, C. Towards low cost automated smartphone- and cloud-based otitis media diagnosis. *Biomedical Signal Processing and Control* 39, 34-52 (2018). https://doi.org/10.1016/j.bspc.2017.07.015
85 Kallio, H., Pietila, A. M., Johnson, M. & Kangasniemi, M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *J Adv Nurs* 72, 2954-2965 (2016). https://doi.org/10.1111/jan.13031
86 Graneheim, U. H. & Lundman, B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Educ Today* 24, 105-112 (2004). https://doi.org/10.1016/j.nedt.2003.10.001
87 Graneheim, U. H., Lindgren, B. M. & Lundman, B. Methodological challenges in qualitative content analysis: A discussion paper. *Nurse Educ Today* 56, 29-34 (2017). https://doi.org/10.1016/j.nedt.2017.06.002
88 Lindgren, B. M., Lundman, B. & Graneheim, U. H. Abstraction and interpretation during the qualitative content analysis process. *Int J Nurs Stud* 108, 103632 (2020). https://doi.org/10.1016/j.ijnurstu.2020.103632
89 Lemkens, N. *et al.* Interpretation of pure-tone thresholds in sensorineural hearing loss (SNHL): a review of measurement variability and age-specific references. *Acta Otorhinolaryngol Belg* 56, 341-352 (2002).
90 Mahomed, F., Swanepoel de, W., Eikelboom, R. H. & Soer, M. Validity of automated threshold audiometry: a systematic review and meta-analysis. *Ear Hear* 34, 745-752 (2013). https://doi.org/10.1097/01.aud.0000436255.53747.a4
91 Oremule, B. *et al.* Mobile audiometry for hearing threshold assessment: A systematic review and meta-analysis. *Clin Otolaryngol* 49, 74-86 (2024). https://doi.org/10.1111/coa.14107
92 Renda, L., Selcuk, O. T., Eyigor, H., Osma, U. & Yilmaz, M. D. Smartphone Based Audiometric Test for Confirming the Level of Hearing; Is It Useable in Underserved Areas? *J Int Adv Otol* 12, 61-66 (2016). https://doi.org/10.5152/iao.2016.1421
93 van Tonder, J., Swanepoel, W., Mahomed-Asmail, F., Myburgh, H. & Eikelboom, R. H. Automated Smartphone Threshold

Audiometry: Validity and Time Efficiency. *J Am Acad Audiol* 28, 200-208 (2017). https://doi.org/10.3766/jaaa.16002

94 Thompson, G. P., Sladen, D. P., Borst, B. J. & Still, O. L. Accuracy of a Tablet Audiometer for Measuring Behavioral Hearing Thresholds in a Clinical Population. *Otolaryngol Head Neck Surg* 153, 838-842 (2015). https://doi.org/10.1177/0194599815593737

95 Saliba, J. *et al.* Accuracy of Mobile-Based Audiometry in the Evaluation of Hearing Loss in Quiet and Noisy Environments. *Otolaryngol Head Neck Surg* 156, 706-711 (2017). https://doi.org/10.1177/0194599816683663

96 Whitton, J. P., Hancock, K. E., Shannon, J. M. & Polley, D. B. Validation of a Self-Administered Audiometry Application: An Equivalence Study. *Laryngoscope* 126, 2382-2388 (2016). https://doi.org/10.1002/lary.25988

97 Corona, A. P., Ferrite, S., Bright, T. & Polack, S. Validity of hearing screening using hearTest smartphone-based audiometry: performance evaluation of different response modes. *Int J Audiol* 59, 666-673 (2020). https://doi.org/10.1080/14992027.2020.1731767

98 Bornman, M., Swanepoel, W., De Jager, L. B. & Eikelboom, R. H. Extended High-Frequency Smartphone Audiometry: Validity and Reliability. *J Am Acad Audiol* 30, 217-226 (2019). https://doi.org/10.3766/jaaa.17111

99 Bright, T. *et al.* Diagnostic accuracy of non-specialist versus specialist health workers in diagnosing hearing loss and ear disease in Malawi. *Trop Med Int Health* 24, 817-828 (2019). https://doi.org/10.1111/tmi.13238

100 Ehlert, K., Heinze, B., Graham, M. A. & Swanepoel, D. Surveillance for ototoxicity in platinum-based chemotherapy using mobile health audiometry with extended high frequencies. *J Laryngol Otol* 137, 61-67 (2023). https://doi.org/10.1017/S0022215122001281

101 Livingstone, D., Talai, A. S., Chau, J. & Forkert, N. D. Building an Otoscopic screening prototype tool using deep learning. *J Otolaryngol Head Neck Surg* 48, 66 (2019). https://doi.org/10.1186/s40463-019-0389-9

102 Zeng, X. *et al.* Efficient and accurate identification of ear diseases using an ensemble deep learning model. *Sci Rep* 11, 10839 (2021). https://doi.org/10.1038/s41598-021-90345-w

103 Khan, M. A. *et al.* Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. *Neural Netw* 126, 384-394 (2020). https://doi.org/10.1016/j.neunet.2020.03.023

104 Lee, J. Y., Choi, S. H. & Chung, J. W. Automated Classification of the Tympanic Membrane Using a Convolutional Neural Network. *Appl Sci-Basel* 9 (2019).
105 Cha, D., Pae, C., Seong, S. B., Choi, J. Y. & Park, H. J. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine* 45, 606-614 (2019). https://doi.org/10.1016/j.ebiom.2019.06.050
106 Chang, K. M. *et al.* Pediatric tympanostomy tube assessment via deep learning. *Am J Otolaryngol* 45, 104334 (2024). https://doi.org/10.1016/j.amjoto.2024.104334
107 Cao, Z. *et al.* Machine Learning in Diagnosing Middle Ear Disorders Using Tympanic Membrane Images: A Meta-Analysis. *Laryngoscope* 133, 732-741 (2023). https://doi.org/10.1002/lary.30291
108 Gibbs, S. Uber's self-driving car saw the pedestrian but didn't swerve – report. *Guardian News & Media Limited* (2018).
109 Viscaino, M. *et al.* Computer-aided diagnosis of external and middle ear conditions: A machine learning approach. *PLoS One* 15, e0229226 (2020). https://doi.org/10.1371/journal.pone.0229226
110 Zafer, C. Fusing fine-tuned deep features for recognizing different tympanic membranes. *Biocybernetics and Biomedical Engineering* 40, 40-51 (2020). https://doi.org/10.1016/j.bbe.2019.11.001
111 Wu, Z. *et al.* Deep Learning for Classification of Pediatric Otitis Media. *Laryngoscope* 131, E2344-E2351 (2021). https://doi.org/10.1002/lary.29302
112 Moberly, A. C. *et al.* Digital otoscopy versus microscopy: How correct and confident are ear experts in their diagnoses? *J Telemed Telecare* 24, 453-459 (2018). https://doi.org/10.1177/1357633X17708531
113 Krause, J. *et al.* Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* 125, 1264-1272 (2018). https://doi.org/10.1016/j.ophtha.2018.01.034
114 Navarathna, N., Kanhere, A., Gomez, C. & Isaiah, A. Artificial intelligence in pediatric otolaryngology: A state-of-the-art review of opportunities and pitfalls. *Int J Pediatr Otorhinolaryngol* 194, 112369 (2025). https://doi.org/10.1016/j.ijporl.2025.112369
115 Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 17, 195 (2019). https://doi.org/10.1186/s12916-019-1426-2

116 Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 3, e745-e750 (2021). https://doi.org/10.1016/S2589-7500(21)00208-9
117 Engstrom, E. & Jebari, K. AI4People or People4AI? On human adaptation to AI at work. *Ai & Society* 38, 967-968 (2022). https://doi.org/10.1007/s00146-022-01464-5
118 Wachter, R. M. & Brynjolfsson, E. Will Generative Artificial Intelligence Deliver on Its Promise in Health Care? *JAMA* 331, 65-69 (2024). https://doi.org/10.1001/jama.2023.25054
119 Furendal, M. & Jebari, K. The Future of Work: Augmentation or Stunting? *Philosophy & Technology* 36, 36 (2023). https://doi.org/10.1007/s13347-023-00631-w
120 Blomgren, K. & Pitkaranta, A. Is it possible to diagnose acute otitis media accurately in primary health care? *Fam Pract* 20, 524-527 (2003). https://doi.org/10.1093/fampra/cmg505
121 Blomgren, K. & Pitkaranta, A. Current challenges in diagnosis of acute otitis media. *Int J Pediatr Otorhinolaryngol* 69, 295-299 (2005). https://doi.org/10.1016/j.ijporl.2004.09.012
122 Jensen, P. M. & Lous, J. Criteria, performance and diagnostic problems in diagnosing acute otitis media. *Fam Pract* 16, 262-268 (1999). https://doi.org/10.1093/fampra/16.3.262
123 Lannon, C., Peterson, L. E. & Goudie, A. Quality measures for the care of children with otitis media with effusion. *Pediatrics* 127, e1490-1497 (2011). https://doi.org/10.1542/peds.2009-3569
124 Sundvall, P. D., Papachristodoulou, C. E. & Nordeman, L. Diagnostic methods for acute otitis media in 1 to 12 year old children: a cross sectional study in primary health care. *BMC Fam Pract* 20, 127 (2019). https://doi.org/10.1186/s12875-019-1018-4
125 Stephens, J. H. *et al.* A multimodal machine learning algorithm improved diagnostic accuracy for otitis media in a school aged Aboriginal population. *J Biomed Inform* 164, 104801 (2025). https://doi.org/10.1016/j.jbi.2025.104801
126 Zhong, Z. *et al.* Artificial intelligence as an auxiliary tool in pediatric otitis media diagnosis. *Int J Pediatr Otorhinolaryngol* 187, 112154 (2024). https://doi.org/10.1016/j.ijporl.2024.112154
127 Binol, H. *et al.* SelectStitch: Automated Frame Segmentation and Stitching to Create Composite Images from Otoscope Video Clips. *Appl Sci-Basel* 10, 5894 (2020).
128 Shaikh, N. *et al.* Development and Validation of an Automated Classifier to Diagnose Acute Otitis Media in Children. *JAMA Pediatr* 178, 401-407 (2024). https://doi.org/10.1001/jamapediatrics.2024.0011

129 Habib, A. R. *et al.* Artificial intelligence to classify ear disease from otoscopy: A systematic review and meta-analysis. *Clin Otolaryngol* 47, 401-413 (2022). https://doi.org/10.1111/coa.13925
130 Ngombu, S., Binol, H., Gurcan, M. N. & Moberly, A. C. Advances in Artificial Intelligence to Diagnose Otitis Media: State of the Art Review. *Otolaryngol Head Neck Surg* 168, 635-642 (2023). https://doi.org/10.1177/01945998221083502
131 van Brink, J. & Gisselsson-Solen, M. Quality of life in Swedish children receiving grommets - An analysis of pre- and postoperative results based on a national quality register. *Int J Pediatr Otorhinolaryngol* 120, 44-50 (2019). https://doi.org/10.1016/j.ijporl.2019.02.009
132 Maile, E. J., Tharu, P. B., Blanchford, H. L., Edmiston, R. & Youngs, R. Quality of life of Nepali patients with ear disease before and after corrective surgery. *Trop Med Int Health* 20, 1041-1047 (2015). https://doi.org/10.1111/tmi.12516
133 Orji, A. *et al.* Global and regional needs, unmet needs and access to hearing aids. *Int J Audiol* 59, 166-172 (2020). https://doi.org/10.1080/14992027.2020.1721577
134 McPherson, B. Hearing assistive technologies in developing countries: background, achievements and challenges. *Disabil Rehabil Assist Technol* 9, 360-364 (2014). https://doi.org/10.3109/17483107.2014.907365
135 Schwalbe, N. & Wahl, B. Artificial intelligence and the future of global health. *Lancet* 395, 1579-1586 (2020). https://doi.org/10.1016/S0140-6736(20)30226-9
136 Mukara, K. B., Waiswa, P., Lilford, R. & Tucci, D. L. Knowledge and care seeking practices for ear infections among parents of under five children in Kigali, Rwanda: a cross-sectional study. *BMC Ear Nose Throat Disord* 17, 7 (2017). https://doi.org/10.1186/s12901-017-0040-1
137 Rupa, V., Jacob, A. & Joseph, A. Chronic suppurative otitis media: prevalence and practices among rural South Indian children. *Int J Pediatr Otorhinolaryngol* 48, 217-221 (1999). https://doi.org/10.1016/s0165-5876(99)00034-8
138 Weir, N. Ear surgery camps in Nepal and the work of the Britain Nepal Otology Service (BRINOS). *J Laryngol Otol* 105, 1113-1115 (1991). https://doi.org/10.1017/s0022215100118377
139 Ernst, E. Ear candles: a triumph of ignorance over science. *J Laryngol Otol* 118, 1-2 (2004). https://doi.org/10.1258/002221504322731529

140 Davis, A. *et al.* Aging and Hearing Health: The Life-course Approach. *Gerontologist* 56 Suppl 2, S256-267 (2016). https://doi.org/10.1093/geront/gnw033
141 David, D. & Werner, P. Stigma regarding hearing loss and hearing aids: A scoping review. *Stigma and Health* 1, 59-71 (2016). https://doi.org/10.1037/sah0000022
142 Bright, T., Mulwafu, W., Thindwa, R., Zuurmond, M. & Polack, S. Reasons for low uptake of referrals to ear and hearing services for children in Malawi. *PLoS One* 12, e0188703 (2017). https://doi.org/10.1371/journal.pone.0188703
143 Van der Meer, G. & Loock, J. W. Why patients miss follow-up appointments: a prospective control-matched study. *East Afr J Public Health* 5, 154-156 (2008). https://doi.org/10.4314/eajph.v5i3.38994
144 Frost, L., Jenkins, L. S. & Emmink, B. Improving access to health care in a rural regional hospital in South Africa: Why do patients miss their appointments? *Afr J Prim Health Care Fam Med* 9, e1-e5 (2017). https://doi.org/10.4102/phcfm.v9i1.1255
145 Diaconu, K. *et al.* Methods for medical device and equipment procurement and prioritization within low- and middle-income countries: findings of a systematic literature review. *Global Health* 13, 59 (2017). https://doi.org/10.1186/s12992-017-0280-2
146 Oh, S. *et al.* Physician Confidence in Artificial Intelligence: An Online Mobile Survey. *Journal of medical Internet research* 21, e12422 (2019). https://doi.org/10.2196/12422
147 Castagno, S. & Khalifa, M. Perceptions of Artificial Intelligence Among Healthcare Staff: A Qualitative Survey Study. *Front Artif Intell* 3, 578983 (2020). https://doi.org/10.3389/frai.2020.578983
148 Scheetz, J. *et al.* A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Sci Rep* 11, 5193 (2021). https://doi.org/10.1038/s41598-021-84698-5
149 Heinrichs, H., Kies, A., Nagel, S. K. & Kiessling, F. Physicians' Attitudes Toward Artificial Intelligence in Medicine: Mixed Methods Survey and Interview Study. *J Med Internet Res* 27, e74187 (2025). https://doi.org/10.2196/74187
150 Mousavi Baigi, S. F., Sarbaz, M., Darroudi, A. & Kimiafar, K. Medical and Paramedical Students' Perspectives on Artificial Intelligence in a Low- and Middle-Income Country: A Cross-Sectional Study. *Health Sci Rep* 8, e71046 (2025). https://doi.org/10.1002/hsr2.71046

151 Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min* 10, 35 (2017). https://doi.org/10.1186/s13040-017-0155-3

152 Charilaou, P. & Battat, R. Machine learning models and over-fitting considerations. *World J Gastroenterol* 28, 605-607 (2022). https://doi.org/10.3748/wjg.v28.i5.605

153 Malterud, K. Qualitative research: standards, challenges, and guidelines. *Lancet* 358, 483-488 (2001). https://doi.org/10.1016/S0140-6736(01)05627-6

154 Hörselskadades riksförbund. Väntetid till hörapparatutprovning. (https://hrf.se/nyheter/upp-till-tva-ars-vantetid-pa-horapparater-omanskligt-och-halsofarligt/, 2025).