Value-based and data-driven vestibular schwannoma care

Neve, O.M.

Value-based and data-driven vestibular schwannoma care

OLAF NEVE

Value-based and data-driven vestibular schwannoma care

Olaf Neve

Design: Studio Winter | www.studiowinter.nl
Printed by: Optima Grafische Communicatie| www.ogc.nl
ISBN: 978-94-6510-245-0
Financial support for publication of this thesis was provided by: MSB GHZ, BeterHoren, Schoonenberg, Chipsoft, Allergy therapeutics, Meditop, emiD

Value-based and data-driven vestibular schwannoma care

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 6 november 2024
klokke 13:00 uur

door

Olaf Maarten Neve
geboren te Hoorn
in 1993

Promotores:	Prof. dr. P.P.G. van Benthem
	Prof. dr. A.M. Stiggelbout
Co-promotor:	Dr. E.F. Hensen
Promotiecommissie:	Prof. dr. J.C. Jansen (LUMC)
	Prof. dr. H.P.M. Kunst (Radboud UMC)
	Prof. dr. S. Verberne (Leiden Institute of Advanced Computer Science)
	Prof. dr. W.J.W. Bos (LUMC)

TABLE OF CONTENTS

CHAPTER 1

General introduction

Vestibular schwannomas are rare benign intracranial tumors. The disease, treatment, and its sequelae impact patients' daily life and the treatment decision-making process is complex. The management of these tumors requires a multidisciplinary approach including specialties such as otorhinolaryngology, neurosurgery, and radiation oncology.

The complex nature of vestibular schwannoma care makes it a suitable subject to reorganize the delivery of care to improve quality, and how to measure and evaluate the quality-of-care delivery. In this thesis, principles of value-based and data-driven healthcare are studied in the context of vestibular schwannoma care.

Value-based healthcare is a management strategy focusing on optimizing value, defined as outcomes relevant to patients divided by the cost necessary to achieve those outcomes. Value creation can be achieved through or supported by data driven technologies that improve accuracy or reduce workload during the health delivery process. In addition, these technologies may support information provision in clinical practice, thereby facilitating improved shared decision making.

In this thesis, several outcomes of vestibular schwannoma care are assessed that are important in understanding the patients' perspective on the disease, the treatment decisions, and the care delivery. In addition, the thesis aims to evaluate the feasibility of tools based on artificial intelligence technologies that facilitate the gathering of essential information for shared treatment decision making and the evaluation of the patients' feedback on the quality of care.

Section 1.1 provides the reader with a background in vestibular schwannoma. Section 1.2 contains a more detailed description of the management strategy of value-based healthcare and section 1.3 the development of data driven healthcare. In section 1.4 the outline of the thesis is described.

1.1. VESTIBULAR SCHWANNOMA

Vestibular schwannomas are rare benign tumors arising from the Schwann cells of the vestibular nerve.[1] The tumor is unilateral in more than 95% of the cases. In a small minority, a genetic disorder (neurofibromatosis type 2 schwannomatosis) can cause bilateral vestibular schwannomas.[2]

Symptoms

Most likely, the first description of vestibular schwannoma was reported in 1777 by Sandifort (1742-1814), a professor of pathology at Leiden University. In a postmortem examination of a woman with single-sided deafness Sandifort noticed and depicted "*an certain hard body attached to the auditory nerve".* [3-5] According to Sandifort "the hard body" was causing hearing loss by affecting the nervous system dedicated to the hearing sense. [3] This accurate observation and clinical reasoning based on the tumor and the adjacent anatomical structures by Sandifort is still relevant and helps to explain and understand symptoms and treatment-associated morbidity of vestibular schwannoma patients. The internal auditory canal contains the cochlear nerve, the superior and inferior vestibular nerve, and the facial nerve. In contrast to the observations of Sandifort, most vestibular schwannomas originate in the inferior vestibular nerve. [6] When tumors are progressive, protrusion towards the cerebellopontine angle occurs. Further progression may eventually lead to compression of the trigeminal nerve and the brainstem.

Progressive audiovestibular complaints are still the most common symptom of vestibular schwannoma patients. Typically patients present with hearing loss (95%), dizziness (61%), and/or tinnitus (63%).[7] However, facial numbness (8%) and headache (12%) may also be presenting symptoms. [7] Symptoms seem to be associated with larger tumor size, especially hearing loss and dizziness. [7-9] Tumor progression and symptoms progression are only weakly associated, so increasing hearing loss does not necessarily indicate tumor progression. [1]

Natural course

The incidence of vestibular schwannoma has increased over the last decades. [10] A large Danish population study showed an increase of the incidence from 3 per million person-years in 1974 to 34 per million person-years in 2015, while at the same time, the average tumor size at diagnosis has decreased from 26mm to 7mm. [10] A Dutch study reported a 16 per million person-years incidence rate, with large regional differences up to 33 per million person-years in some regions. [11] The rising incidence combined with the regional differences and the smaller average tumor size strongly suggest that increased availability of diagnostic MRI leads to more frequent detection of smaller vestibular schwannoma. However, additional yet unknown etiologies may also attribute to the increasing incidence. [12] The shift in tumor size distribution also impacts vestibular schwannoma management. In smaller tumors, active surveillance is nowadays increasingly advocated at the expense of surgery or radiotherapy as management strategy of choice at initial presentation. [13, 14]

This shift in treatment strategy found its origin in the increased insight into the natural course of sporadic vestibular schwannoma. After 10 years follow-up, 42-46% of the tumors show progression and a small majority remain indolent. [15-17] Tumor size at baseline is currently the only identified predictor of tumor progression. [18-20] Tumor progression is pivotal in deciding between treatment strategies. All treatment strategies aim to prevent future serious complications due to progressive tumors, such as brain stem compression and increased intracranial pressure. None of the currently available treatments can eradicate the tumor without risk of damage to the audiovestibular organs or nerves, and therefore active treatment is generally not aimed at alleviating the audiovestibular symptoms most patients present with. Indeed, hearing and dizziness outcomes are poor across treatment modalities and active treatment (i.e., surgery or radiotherapy) yields additional risks, such as that of facial paresis. [1, 21, 22] Therefore, maintaining quality of life is an important aim of vestibular schwannoma management, additional to tumor control.

Treatment strategy

The treatment option of choice for indolent tumors is active surveillance, during which tumor progression is monitored using MR imaging. [23] Initially, tumors are monitored annually or even after six months on gadolinium-enhanced T1 or high-resolution T2 MRI. When there is no tumor progression on several consecutive scans the time interval is prolonged. Tumor progression may occur, albeit infrequently, even after prolonged observation without progression. [15] Therefore, lifelong surveillance is advised. [24] Tumor progression on two consecutive MRIs is commonly defined as a >2mm increase in diameter. This limit is based on the human measurement error of the tumor diameter. [25, 26] Progression occurs more often in cystic tumors, and tumor growth within the first year of active surveillance is an indicator of future progression. [17, 18] During active surveillance, deterioration of hearing is common even in stable tumors. [27, 28] The unclear relation between tumor progression and increasing audiovestibular symptoms highlights the need for periodical MRI examination. In case of tumor progression, active treatment such as surgery or radiotherapy is advised, although continuing active surveillance might also be reasonable, especially in smaller tumors. [1, 23]

Surgery has historically been the only active treatment for vestibular schwannoma. The aim of surgery is to achieve tumor control while maintaining functionality. In the late nineteenth century surgery was rather hazardous with mortality rates over 75%.[29] During the twentieth-century, surgical techniques improved, as did perioperative care, resulting in mortality rates <1% and excellent tumor control. [4] Surgery is the preferred treatment option for very large tumors that compress the brainstem. [23, 30] In smaller tumors, surgery can be a suitable treatment option too, and smaller tumors have lower

postoperative complication rates than larger tumors. [1] The translabyrinthine approach, which inherently leads to hearing loss since the inner ear is sacrificed, results in tumor control in 94% of patients and facial nerve preservation in 85% of patients. [31] The retro sigmoidal approach can preserve some degree of hearing in 35% of the cases, and has a reported tumor control over 95% of patients and comparable facial nerve preservation. [32, 33]

Since 1995 radiotherapy has increasingly been performed in vestibular schwannoma patients. [1, 4] The aim of radiotherapy is to stop tumor progression and preserve neurological function, however the tumor is not eradicated. Stereotactic radiotherapy is offered to patients with tumors smaller than 30 mm extrameatal diameter. [1] Preferably, radiotherapy is chosen in tumors smaller than 25mm maximal extrameatal diameter to prevent complications of radiation-induced pseudoprogression, which can occur in the first two years after radiotherapy. [34, 35] Various types of fractioning are used, ranging from single fraction to 28 fractions with similar tumor control rates of 91-94%.[36-38] Hearing preservation in the long term is <50%, but facial nerve preservation is above 95%.[37] The risk of radiotherapy-induced malignancies is negligibly small in sporadic cases. [39]

When tumor control is not established by either surgery or radiotherapy, salvage therapy is indicated. Salvage therapy could be surgery after initial radiotherapy or vice versa. [1, 23] Salvage therapy yields a higher risk of treatment-associated complications and lower neurological functionality. [30]

Decision making

Both active treatment strategies have completely different delivery methods (e.g., inpatient vs. outpatient, invasive vs. non-invasive), but comparable outcomes. Therefore, the patients' perspectives and preferences play a central role in decision-making. This is reflected in the interest in quality of life as subsidiary outcome in vestibular schwannoma care. [40-43] In the short term, quality of life seems to be affected by the diagnosis rather than the treatment strategies. [41] Tinnitus and dizziness seem to be the symptoms that have the largest impact on quality of life. [40, 44] Both symptoms are not likely to improve after any of the treatment strategies. There is a delicate tradeoff between current symptoms, future complications, and potential side effects of active treatment. Patients are captured in a balancing act in the decision-making process, as they know that they may grow old with the tumor and its sequalae and that the aim of treatment is to prevent future complications. These complications can be devastating for quality of life, but active treatment will in the short term most likely worsen symptoms rather than relieve them.

The complex clinical decision-making process in vestibular schwannoma care warrants a shared decision making. Shared decision making entails a close collaboration between patients and clinician in which they explore the available options, and look for the best choice considering both the patient's and the medical context. [45]

For the different treatment options different medical specialties are involved. To optimize the information provision for patient, vestibular schwannoma care should be organized around the disease, including all involved healthcare professionals. Such an organization will probably reduce the known healthcare provider-driven demand, in which the expertise and experience of the physician rather than the preference of the patient determine treatment. [46]

Vestibular schwannoma care at Leiden University Medical Center

Leiden University Medical Center (LUMC) is a national expert center for vestibular schwannoma. Vestibular schwannoma care is organized in a multidisciplinary care pathway. In 1995 the pathway was started with a biweekly multidisciplinary team meeting including otorhinolaryngologists, neurosurgeons, radiation oncologists and radiologists. Over the years, this has led the LUMC becoming an expert center in which over 1200 vestibular schwannoma patients are discussed annually in the weekly meetings. Each year, 200-260 new patients visit the expert center; most of them are diagnosed at another hospital and referred for counseling and management.

At the initial consultation with a neurotologist or neurosurgeon, the symptoms are registered, and information about the disease and its natural course is provided. Every patient is discussed in the multidisciplinary team meeting to check the indication for active treatment and possible contra-indications for treatment options. When active treatment is indicated, the patient is invited to an interdisciplinary outpatient clinic to discuss the viable treatment options in two consecutive consultations with a radiation oncologist and a surgeon (either neurosurgeon or otorhinolaryngologist). Together, the patient and the physician decide to pursue one of the three main treatment strategies. Surgery is performed at the LUMC, radiotherapy is performed either at the LUMC or at a radiotherapy facility closer to the patients' residency. Active surveillance using MRI scans, as the MRI scan for follow-up after therapy can also be acquired at the referring hospital.

At an organizational level, the care process, outcomes, and improvement trajectories are monitored and discussed in monthly management meetings. The management team consists of one representative of each specialty (otorhinolaryngology, neurosurgery, and radiation oncology), the case manager (a nurse), a data scientist, and

supportive staff of the quality and safety department. In 2018, the management team developed a core outcome set to be measured in each patient, consisting of several clinical parameters, such as symptoms, hearing loss, facial nerve function, etc. These outcomes are noted in the electronic patient records in structured fields. The outcomes at group level are summarized in dedicated dashboards. This information helps the management team to monitor quality and evaluate organizational improvements. Since 2019 vestibular schwannoma care has been reorganized according to the value based healthcare principles.

1.2. VALUE-BASED HEALTHCARE

In 1966 Donabedian formulated the foundations of quality measurement in healthcare by addressing the three components (structure, process, and outcome) that all contribute to quality of care. [47, 48] According to Donabedian, all three categories should be assessed when evaluating the quality of care. [47, 49] Structure encompasses the setting in which healthcare delivery occurs. The assessment of structure comprises the organization of care, the equipment, and the qualifications of staff. Given a good structure, it is more likely that the process of care will be adequate. This process of care is about what is done in the healthcare delivery and whether that is considered as good medical care according to scientific medical associations. Assessing this process is based on appropriateness, completeness, and the lack of redundancy of the actual healthcare delivery. Process assessment relies on standards, protocols, and compliance. When the process is good, the likelihood of a better outcome is higher. Outcome comprises the effect of healthcare delivery on the patient's health status and satisfaction. Outcomes should be assessed with clearly defined and relevant measurement tools. Outcomes of care are the most relevant component when measuring quality of care, however, outcomes can be influenced by other factors than quality of care, and sometimes outcomes only become apparent after long periods of time. [47, 49] Both these aspects complicate the measurement of healthcare quality through the assessment of outcomes.

Since Donabedian initiated the quality of care movement, continuous quality measurements guiding quality improvements have been increasingly used in clinical practice. [50] At the beginning of this century in 2001, the Institute of Medicine (IOM) published the influential report *Crossing the Quality Chasm.* [48, 51] The report urged to change the healthcare delivery in the United States in order to increase the quality of care in the 10 years to follow. At that moment, the healthcare system was deemed unfit for healthcare challenges in the 21st century due to the upcoming shift from acute to chronic care, an increasingly aging population, and quickly rising healthcare costs. The IOM formu-

lated the fundamentals for 21st-century quality of care to reshape the system. These fundamentals noted that healthcare should be safe, effective, patient-centered, timely, efficient, and equitable. [51]

In 2006, Harvard business school professors Michael Porter and Elisabeth Teisberg noticed slow progress in performance improvement, due to conflicting interests of stakeholders in healthcare. Their solution was to develop an overarching goal for all healthcare stakeholders: increasing patient value. They defined value as outcomes relevant to patients divided by the cost needed to achieve them. They stated that when patient value improves, patients and providers will both benefit, as will payers. [52, 53] Measurement and improvement of value can drive system progress. Competition of healthcare providers maximizing patient value will result in system broad performance improvement. The concept of value-based healthcare(VBHC) was born.

Value-based healthcare components

Porter proposed six components that were necessary to deliver VBHC. [52, 54] First, healthcare delivery should be organized around a disease in an integrated practice unit (IPU). An IPU is a team of dedicated healthcare professionals who treat a specific disease or medical condition and its sequelae. In Porter's vision, the aim of an IPU should be to maximize patient value. The IPU is responsible for the full cycle of care and the IPU-members should meet regularly to discuss patients as well as the care process and outcomes.

Second, outcomes and costs should be measured for every patient in the VBHC management strategy, to measure and quantify improvements. According to the VBHC principles, it is essential that the outcomes are relevant to patients and cover the complete cycle of care. Porter describes three levels of outcomes: health status achieved (i.e., mortality and functional status), process of recovery (i.e., time to recovery and disutility of care/treatment), and sustainability of health (i.e., absence of recurrences and long-term consequences). [53] In addition to patient outcomes, healthcare delivery costs should be measured. Measuring both outcomes and costs helps to steer improvements to enhance patient value.

Third, reimbursement should move from fee-for-service to bundled payments for care cycles. Reimbursement systems can influence care delivery. Fee for service, the dominant system in the United States at the moment that Porter and Teisberg developed VBHC, motivates healthcare providers to increase their production, which does not necessarily generate more value. Bundled payments for the complete cycle of care on

the other hand will stimulate providers to achieve efficient healthcare delivery while maintaining or improving outcomes.

Fourth, care delivery should be integrated across separate facilities. Healthcare delivery is siloed in different organizations, such as hospitals and primary care practices. Each hospital provides a large variety of care services. Porter and Teisberg argue for concentration of care and differentiation of organizations. Instead of diversifying their efforts across numerous areas, hospitals should concentrate their expertise on a select few conditions, as the authors claim that specialization helps to optimize patient value.

Fifth, excellent services should be expanded across geography. Healthcare delivery is organized locally, even large academic centers have a locally oriented adherence area. When IPUs deliver excellent care, they should expand their geographical adherence using an affiliation network. Low complexity and low-cost diagnostic and treatment services should be provided regionally, and high complexity, high-cost services in a few dedicated centers. This geographical expansion will improve efficiency and lead to better value.

Sixth, an enabling information technology platform should assist the previous components. A supporting IT platform is essential to achieve better value. Such an IT platform should be easily accessible, patient-centered, and encompass the complete care cycle. In addition, data extraction and data portability should be possible to track outcomes and costs.

Value-based healthcare in Europe

Over the years, elements of VBHC have been adopted and implemented as a strategy for healthcare delivery in various countries in different continents. [55, 56] However, VBHC, as proposed by Porter, requires a radical change in healthcare delivery and a reshape of healthcare organizations. Adoption of VBHC is not always compatible with national and local health policies and can conflict with local governance structures. [56] As a result, various VHBC variants have been conceptualized. In most cases, several components of Porter and Teisberg's agenda have been adapted. [55-59] Other components have been omitted and sometimes new components have been included in the VBHC concept.

In European countries, VBHC does not have the primary focus on value-based competition as described in the original concept of Porter and Teisberg. [52, 55, 60, 61] Across Europe, there has been more emphasis on measuring health outcomes and patient-centeredness. [61] For example, in Sweden measuring outcomes and increasing the patients' perspective have been embraced while other aspects are less well adopted. [62, 63] In the Netherlands,

the conceptualization of VBHC has mainly focused on the components 'organizing care in IPUs' and 'measuring outcomes relevant to patients'.[64] Measuring healthcare costs, geographical expansion, and bundled payments play a less prominent role.[65] In addition, shared decision-making is seen as an important aspect of VBHC in the Netherlands, while it was not mentioned in the original work on Porter's VBHC concept.[64, 66]

Value-based healthcare in the Netherlands

This thesis focusses on aspects of the VBHC model as it has conceptualized in the Netherlands. This Dutch version of VBHC aims to provide continuous quality improvement and empower patients to achieve true patient-centered care delivery.[64] Prominent components of this Dutch VBHC version are 'organizing care in IPUs', measuring patient outcomes on an individual and a group level, shared decision making, and an enabling data platform.

Shared decision-making is seen by healthcare professionals and patient associations in the Netherlands as an essential aspect to increase the patient-centeredness of care.[64] Incorporating shared decision-making in VBHC increases the focus on value creation for individual patients. Shared decision-making is expected to be improved by using patient-reported outcomes. These can help to prioritize patient problems and to improve communication between doctors and patients.[67, 68]

In addition to value creation at the individual level, measuring patient outcomes structurally help IPUs with continuous quality improvement.[68] A use case of a breast cancer IPU reported that organizing care around patients and incorporating structural outcome measurements resulted in better insight in care delivery and opportunities for improvement.[68] Furthermore, measurements can be used to benchmark outcomes between hospitals, as is shown in an example for cardiac disease care in the Netherlands.[69] This initiative has measured outcomes of fourteen heart centers in the Netherland, which were then used to benchmark, improve and learn from the comparisons. The program resulted in (slightly) better patient outcomes and better patient satisfaction.[70, 71]

An important factor in VBHC implementation is a data platform that links clinical and patient information of several sources and provides overviews of outcomes on both an individual and IPU level. Analyzing, understanding, and visualizing these large amounts of complex data is laborious and requires data handling skills. Data-driven techniques might contribute to this process and help to achieve better VBHC delivery.

1.3. DATA-DRIVEN HEALTHCARE

As argued above, VBHC implies measuring outcomes and costs of every patient. These outcomes can be collected using different data sources, such as electronic health records (EHRs), PROM collecting platforms, and radiological picture archiving and communication systems (PACS). Transforming these data into information that can be used in the existing clinical and organizational workflow is key in creating value. Retrieving valuable information from large, various and unstructured data is a challenging topic within the field of data science. [72]

The use of data science techniques in healthcare to improve care delivery has been called data-driven healthcare. In this thesis, two specific aspects of data-driven healthcare are highlighted. First, the provision of real-time, context-dependent and actionable information in the clinical workflow of physicians as a means to contribute to patient value. [73] Second, the use of data-driven technologies such as artificial intelligence techniques as a means to improve the quality of care and/or reduce clinical workload. [74]

The first aspect involves combining data sources and retrieving unstructured data. Both processes provide real time insight at the individual patient and IPU levels, and require advanced data analytics and management technologies. [74] To overcome these challenges, collaboration between technical and software experts and clinicians is required. The use of PROMs in clinical practice during consultations and group level data for improvement at the IPU level can only be established when clinicians can seamlessly integrate the right information at the right moment in their clinical workflow. [75]

Linking and summarizing data from multiple sources is challenging. [67, 70, 76] The introduction of EHRs has led to a rapidly increasing amount of data collected per patient. [72] Much of the information of EHRs is entered in unstructured free text fields making automated data extraction and analysis complex. [76] Increasing the structured input in EHR fields is one of the options to facilitate retrieval of clinical data. Furthermore, selecting the right patient groups and summarizing the core outcomes set using clinical data from different sources is essential. [75] These sources can be the EHR, PACS, and PROM outcomes using special questionnaire distribution software. More hospitals have built so-called ‘data lakes’ to store, label, and access automatically all raw data acquired by various systems in their organization. Such a data lake can improve the accessibility and reusability of data. [77] Decisions about the way in which clinically relevant data from this data lake are selected and visualized are essential for their usability, and require the involvement of clinicians, as the end-users of the dashboard.

The second aspect of data-driven healthcare concerns software that performs specific clinical tasks or assists in the analysis of large amounts of data.[74] Often the software uses artificial intelligence (AI) techniques to recognize certain patterns, for example to detect a specific pathology on an X-ray. Theoretically, these software systems can reduce the time and workload. Sometimes the systems can even outperform clinicians and reduce medical errors. In both ways the software contributes to increasing patient value.[78, 79] Three major fields of AI applications can be described in healthcare. First, clinicians can be supported in diagnosis by rapidly and accurately detecting pathology on imaging. Several algorithms in the field of radiology, pathology, dermatology, gastroenterology, and cardiology have been trained to detect abnormalities on imaging, pathology slices, clinical photographs, coloscopies, or electrocardiograms. [74] Second, algorithms can be trained to predict clinical outcomes such as mortality or readmission. These predictions can help clinicians to make treatment decisions. [74] Last, AI can help to analyze large amounts of data in research settings such as genome studies, neuroanatomy or laboratory tests. [74] Although the use AI-based health care tools seems promising, the implementation of advanced data-driven technology in clinical practice is still limited. [74, 78]

The lack of adoption into clinical practice thus far can be explained by difficulties in the integration into the complex framework of clinical practice, and clinicians that are not used to or trained to work with the new tools. [80] Furthermore, the quality of the algorithms relies fully on the quality of the data they are trained on. Many datasets have poor quality or are context-dependent, or algorithms and tools based on such datasets may not be applicable in different clinical contexts. [81]

In this thesis, principles of value-based and data-driven healthcare are studied in the context of vestibular schwannoma care.

1.4. THESIS OUTLINE

This thesis consists of two sections. The first section on value-based vestibular schwannoma care, assessed several outcomes of vestibular schwannoma care that are important in understanding the patients' perspective on the disease, the treatment decisions, and the care delivery. The second section on data-driven vestibular schannoma care evaluated the feasibility of tools based on artificial intelligence technologies that facilitate the gathering of essential information for shared treatment decision making and the evaluation of the patients' feedback on the quality of care.

Chapters 2 and 3 provide insights into the long-term quality of life and employment status of vestibular schwannoma patients. Quality of life of life and employment status are important and relevant outcomes of any disease or intervention as they illustrate the ability of a patient to function in daily life and as a member of society. These important outcomes are used to evaluate quality of care at the group level and improve shared decision-making for individual patients, and as such contribute to VBHC in the vestibular schwannoma care pathway.

Patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) are essential to measure and improve outcomes and experiences relevant to patients. Most PROMs and PREMs are questionnaires, which only provide helpful information when patients complete and return them. High response rates reduce the likelihood of response bias and increase the reliability of the outcomes. When evaluating any patient reported measure, optimizing patient response rates is therefore important. In **chapter 4**, the effect of the delivery method on the response rate of these questionnaires is evaluated.

In **chapter 5**, patient factors that influence the shared decision-making process are discussed. This provides insight into what drives patients when making important treatment decisions together with their physician. Understanding these factors may improve proper shared decision-making as it sheds light on the patients' perspective.

A novel methodology to measure and analyze patient experiences is described in **chapters 6 and 7**. The answers of patients to open-ended questions about their experience in the vestibular schwannoma care pathway reflect the patient experience more comprehensively compared to classical close-ended PREMs and allow for specific interventions to improve this experience. The use of AI techniques reduces the workload associated with evaluating the answers to open-ended questions, thereby facilitating implementation into clinical practice and increasing the added value of the open-ended PREM.

Chapters 8 and 9 describe a new method to measure tumor volume efficiently and accurately using an automated tool. Determining tumor progression is essential in the clinical decision-making for vestibular schwannoma patients. However, diameter measurement, which is currently the golden standard[82], is known to have a large inter- and intraobserver variability. Volume measurement is more accurate but also very time-consuming. [26] An automated tool measuring both might be the solution to improve the quality of measurements while reducing the workload for clinicians.

In **chapter 10** the results and implications of the previous chapters for the value-based and data-driven vestibular schwannoma care are discussed and suggestions for future research are provided.

REFERENCES

1. Carlson ML, Link MJ. Vestibular Schwannomas. *N Engl J Med*. 2021;384(14):1335-1348. doi:10.1056/nejmra2020394
2. Tamura R. Current Understanding of Neurofibromatosis Type 1, 2, and Schwannomatosis. *Int J Mol Sci*. 2021;22(11):5850. doi:10.3390/ijms22115850
3. Sandifort E. *De duram quodam corpusculo, nervo auditorio adherente; observationes anatomico-pathologicae*. vol Book 1. 1777.
4. Ramsden RT. The bloody angle: 100 years of acoustic neuroma surgery. *J R Soc Med*. 1995;88(8):464P-468P.
5. Huang AE, Marinelli JP, Link MJ, Boes CJ, Carlson ML. A Journey Through 100 Years of Vestibular Schwannoma Surgery at Mayo Clinic: A Historical Illustrative Case Series. *Otol Neurotol*. 2020;41(10):e1379-e1392. doi:10.1097/mao.0000000000002888
6. Khrais T, Romano G, Sanna M. Nerve origin of vestibular schwannoma: a prospective study. *The Journal of Laryngology & Otology*. 2008;122(2):128-131. doi:10.1017/s0022215107001028
7. Matthies C, Samii M. Management of 1000 vestibular schwannomas (acoustic neuromas): clinical presentation. *Neurosurgery*. 1997;40:1-9; discussion 9-10.
8. Harun A, Agrawal Y, Tan M, Niparko JK, Francis HW. Sex and Age Associations With Vestibular Schwannoma Size and Presenting Symptoms. *Otol Neurotol*. 2012;33(9):1604-1610. doi:10.1097/MAO.0b013e31826dba9e
9. Patel NS, Huang AE, Dowling EM, et al. The Influence of Vestibular Schwannoma Tumor Volume and Growth on Hearing Loss. *Otolaryngology–Head and Neck Surgery*. 2020;162(4):530-537. doi:10.1177/0194599819900396
10. Reznitsky M, Petersen MMBS, West N, Stangerup S-E, Cayé-Thomasen P. Epidemiology Of Vestibular Schwannomas – Prospective 40-Year Data From An Unselected National Cohort</p>. *Clin Epidemiol*. 2019;Volume 11:981-986. doi:10.2147/clep.s218670
11. Kleijwegt M, Ho Y, Visser Y, Godefroy W, Van Der Mey A. Real Incidence of Vestibular Schwannoma? Estimations From a National Registry. *Otol Neurotol*. 2016;37:1411-1417. doi:10.1097/MAO.0000000000001169
12. Marinelli JP, Lohse CM, Grossardt BR, Lane JI, Carlson ML. Rising Incidence of Sporadic Vestibular Schwannoma: True Biological Shift Versus Simply Greater Detection. *Otol Neurotol*. 2020;41(6):813-847. doi:10.1097/mao.0000000000002626
13. Chan SA, Marinelli JP, Hahs-Vaughn DL, Nye C, Link MJ, Carlson ML. Evolution in Management Trends of Sporadic Vestibular Schwannoma in the United States Over the Last Half-century. *Otol Neurotol*. 2021;42(2):300-305. doi:10.1097/mao.0000000000002891
14. Torres Maldonado S, Naples JG, Fathy R, et al. Recent Trends in Vestibular Schwannoma Management: An 11-Year Analysis of the National Cancer Database. *Otolaryngology–Head and Neck Surgery*. 2019;161(1):137-143. doi:10.1177/0194599819835495
15. Reznitsky M, Petersen MMBS, West N, Stangerup S-E, Cayé-Thomasen P. The natural history of vestibular schwannoma growth—prospective 40-year data from an unselected national cohort. *Neuro Oncol*. 2021;23(5):827-836. doi:10.1093/neuonc/noaa230
16. Yoshimoto Y. Systematic review of the natural history of vestibular schwannoma. *J Neurosurg*. 2005;103(1):59-63. doi:10.3171/jns.2005.103.1.0059
17. Paldor I, Chen AS, Kaye AH. Growth rate of vestibular schwannoma. *J Clin Neurosci*. 2016/10/01/2016;32:1-8. doi:https://doi.org/10.1016/j.jocn.2016.05.003

18. Kleijwegt M, Bettink F, Malessy M, Putter H, Van Der Mey A. Clinical Predictors Leading to Change of Initial Conservative Treatment of 836 Vestibular Schwannomas. *Journal of Neurological Surgery Part B: Skull Base*. 2020;81(01):015-021. doi:10.1055/s-0039-1678708
19. D'Haese S, Parmentier H, Keppler H, et al. Vestibular schwannoma: natural growth and possible predictive factors. *Acta Otolaryngol*. 2019;139(9):753-758. doi:10.1080/00016489.2019.1635268
20. Hunter JB, Francis DO, O'Connell BP, et al. Single Institutional Experience With Observing 564 Vestibular Schwannomas. *Otol Neurotol*. 2016;37(10):1630-1636. doi:10.1097/mao.0000000000001219
21. Tveiten OV, Carlson ML, Goplen F, Vassbotn F, Link MJ, Lund-Johansen M. Long-term Auditory Symptoms in Patients With Sporadic Vestibular Schwannoma. *Neurosurgery*. 2015;77(2):218-227. doi:10.1227/neu.0000000000000760
22. Khandalavala KR, Saba ES, Kocharyan A, et al. Hearing Preservation in Observed Sporadic Vestibular Schwannoma: A Systematic Review. *Otol Neurotol*. 2022;doi:10.1097/mao.0000000000003520
23. Goldbrunner R, Weller M, Regis J, et al. EANO guideline on the diagnosis and treatment of vestibular schwannoma. *Neuro Oncol*. 2020;22(1):31-45. doi:10.1093/neuonc/noz153
24. Macielak RJ, Patel NS, Lees KA, et al. Delayed Tumor Growth in Vestibular Schwannoma: An Argument for Lifelong Surveillance. *Otol Neurotol*. 2019;40(9):1224-1229. doi:10.1097/mao.0000000000002337
25. Dunn IF, Bi WL, Mukundan S, et al. Congress of Neurological Surgeons Systematic Review and Evidence-Based Guidelines on the Role of Imaging in the Diagnosis and Management of Patients With Vestibular Schwannomas. *Neurosurgery*. 2018;82(2):E32-E34. doi:10.1093/neuros/nyx510
26. van de Langenberg R, de Bondt BJ, Nelemans PJ, Baumert BG, Stokroos RJ. Follow-up assessment of vestibular schwannomas: volume quantification versus two-dimensional measurements. *Neuroradiology*. Aug 2009;51(8):517-24. doi:10.1007/s00234-009-0529-4
27. Stangerup SE, Tos M, Thomsen J, Caye-Thomasen P. Hearing outcomes of vestibular schwannoma patients managed with 'wait and scan': predictive value of hearing level at diagnosis. *The Journal of Laryngology & Otology*. 2010;124(5):490-494. doi:10.1017/s0022215109992611
28. Stangerup SE, Caye-Thomasen P, Tos M, Thomsen J. Change in hearing during 'wait and scan' management of patients with vestibular schwannoma. *The Journal of Laryngology & Otology*. 2008;122(7):673-681. doi:10.1017/s0022215107001077
29. Ramsden RT. 'A brilliant surgical result, the first recorded': Annandale's case, 3 May 1895. *The Journal of Laryngology & Otology*. 1995;109(5):369-373. doi:10.1017/s0022215100130221
30. Hadjipanayis CG, Carlson ML, Link MJ, et al. Congress of Neurological Surgeons Systematic Review and Evidence-Based Guidelines on Surgical Resection for the Treatment of Patients With Vestibular Schwannomas. *Neurosurgery*. 2018;82(2):E40-E43. doi:10.1093/neuros/nyx512
31. de Boer NP, Koot RW, Jansen JC, et al. Prognostic Factors for the Outcome of Translabyrinthine Surgery for Vestibular Schwannomas. *Otol Neurotol*. 2021;42(3):475-482. doi:10.1097/mao.0000000000002980
32. Preet K, Ong V, Sheppard JP, et al. Postoperative Hearing Preservation in Patients Undergoing Retrosigmoid Craniotomy for Resection of Vestibular Schwannomas: A Systematic Review of 2034 Patients. *Neurosurgery*. 2019;doi:10.1093/neuros/nyz147
33. Ahmad RARL, Sivalingam S, Topsakal V, Russo A, Taibah A, Sanna M. Rate of Recurrent Vestibular Schwannoma after Total Removal via Different Surgical Approaches. *Ann Otol Rhinol Laryngol*. 2012;121(3):156-161. doi:10.1177/000348941212100303

34. Bailo M, Boari N, Franzin A, et al. Gamma Knife Radiosurgery as Primary Treatment for Large Vestibular Schwannomas: Clinical Results at Long-Term Follow-Up in a Series of 59 Patients. *World Neurosurg*. 2016;95:487-501. doi:10.1016/j.wneu.2016.07.117
35. Breshears JD, Chang J, Molinaro AM, et al. Temporal Dynamics of Pseudoprogression After Gamma Knife Radiosurgery for Vestibular Schwannomas—A Retrospective Volumetric Study. *Neurosurgery*. 2019;84(1):123-131. doi:10.1093/neuros/nyy019
36. Soltys SG, Milano MT, Xue J, et al. Stereotactic Radiosurgery for Vestibular Schwannomas: Tumor Control Probability Analyses and Recommended Reporting Standards. *International Journal of Radiation Oncology*Biology*Physics*. 2021;110(1):100-111. doi:10.1016/j.ijrobp.2020.11.019
37. Johnson S, Kano H, Faramand A, et al. Long term results of primary radiosurgery for vestibular schwannomas. *J Neurooncol*. 2019;145(2):247-255. doi:10.1007/s11060-019-03290-0
38. Windisch P, Tonn J-C, Fürweger C, et al. Longitudinal Changes of Quality of Life and Hearing Following Radiosurgery for Vestibular Schwannoma. *Cancers (Basel)*. 2021;13(6):1315. doi:10.3390/cancers13061315
39. Seferis C, Torrens M, Paraskevopoulou C, Psichidis G. Malignant transformation in vestibular schwannoma: report of a single case, literature search, and debate. *J Neurosurg*. 2014;121(Suppl_2):160-166. doi:10.3171/2014.7.gks141311
40. Soulier G, Van Leeuwen BM, Putter H, et al. Quality of Life in 807 Patients with Vestibular Schwannoma: Comparing Treatment Modalities. *Otolaryngology–Head and Neck Surgery*. 2017;157(1):92-98. doi:10.1177/0194599817695800
41. Carlson ML, Tveiten OV, Driscoll CL, et al. Long-term quality of life in patients with vestibular schwannoma: an international multicenter cross-sectional study comparing microsurgery, stereotactic radiosurgery, observation, and nontumor controls. *J Neurosurg*. 2015;122:833-842. doi:10.3171/2014.11.JNS14594
42. Carlson ML, Tombers NM, Kerezoudis P, Celda MP, Lohse CM, Link MJ. Quality of Life Within the First 6 Months of Vestibular Schwannoma Diagnosis With Implications for Patient Counseling. *Otol Neurotol*. 2018;39(10):e1129-e1136. doi:10.1097/mao.0000000000001999
43. Robinett ZN, Walz PC, Miles-Markley B, Moberly AC, Welling DB. Comparison of Long-term Quality-of-Life Outcomes in Vestibular Schwannoma Patients. *Otolaryngology–Head and Neck Surgery*. 2014;150(6):1024-1032. doi:10.1177/0194599814524531
44. Carlson ML, Tveiten ØV, Driscoll CL, et al. What drives quality of life in patients with sporadic vestibular schwannoma? *The Laryngoscope*. 2015;125(7):1697-1702. doi:10.1002/lary.25110
45. Stiggelbout AM, Van der Weijden T, De Wit MPT, et al. Shared decision making: Really putting patients at the centre of healthcare. *BMJ (Clinical research ed)*. 2012;344:e256. doi:10.1136/bmj.e256
46. Carlson ML, Glasgow AE, Grossardt BR, Habermann EB, Link MJ. Does where you live influence how your vestibular schwannoma is managed? Examining geographical differences in vestibular schwannoma treatment across the United States. *J Neurooncol*. 2016;129(2):269-279. doi:10.1007/s11060-016-2170-5
47. Donabedian A. Evaluating the Quality of Medical Care. *Milbank Q*. 1966;44(3):166-203. doi:10.1111/j.1468-0009.2005.00397.x
48. Ayanian JZ, Markel H. Donabedian's Lasting Framework for Health Care Quality. *N Engl J Med*. 2016;375(3):205-207. doi:10.1056/NEJMp1605101
49. Donabedian a. The quality of care. How can it be assessed? *JAMA : the journal of the American Medical Association*. 1988;260:1743-1748. doi:10.1001/jama.260.12.1743

50. Berwick D, Fox DM. “Evaluating the Quality of Medical Care”: Donabedian’s Classic Article 50 Years Later. *The Milbank Quarterly*. 2016;94(2):237-241. doi:10.1111/1468-0009.12189
51. Institute of Medicine Committee on Quality of Health Care in A. *Crossing the Quality Chasm: A New Health System for the 21st Century*. National Academies Press (US) Copyright 2001 by the National Academy of Sciences. All rights reserved.; 2001.
52. Porter ME, Teisberg EO. *Redefining Health Care: Creating Value-based Competition on Results*. Harvard Business School Press; 2006.
53. Porter ME. What is value in health care? *N Engl J Med*. Dec 23 2010;363(26):2477-81. doi:10.1056/NEJMp1011024
54. Porter ME, Lee TH. The Strategy That Will Fix Health Care. *Harv Bus Rev*. 2013;
55. Mjåset C, Ikram B, Nagra NS, Feeley TW. Value-Based Health Care in Four Different Health Care Systems. *NEJM Catalyst*. 2020;doi:10.1056/cat.20.0530
56. Ramsdal H, Bjørkquist C. Value-based innovations in a Norwegian hospital: from conceptualization to implementation. *Public Management Review*. 2020;22(11):1717-1738. doi:10.1080/14719037.2019.1648695
57. Van Staalduinen DJ, Van Den Bekerom P, Groeneveld S, Kidanemariam M, Stiggelbout AM, Van Den Akker-Van Marle ME. The implementation of value-based healthcare: a scoping review. *BMC Health Serv Res*. 2022;22(1)doi:10.1186/s12913-022-07489-2
58. Bonde M, Bossen C, Danholt P. Translating value-based health care: an experiment into healthcare governance and dialogical accountability. *Sociol Health Illn*. 2018;40(7):1113-1126. doi:10.1111/1467-9566.12745
59. Steinmann G, Daniels K, Mieris F, Delnoij D, van de Bovenkamp H, van der Nat P. Redesigning value-based hospital structures: a qualitative study on value-based health care in the Netherlands. *BMC Health Serv Res*. Sep 22 2022;22(1):1193. doi:10.1186/s12913-022-08564-4
60. Cossio-Gil Y, Stamm T, Omara M, et al. The roadmap for implementing value based healthcare in European university hospitals - consensus report and recommendations. *medRxiv*. 2021:2021.05.18.21257238. doi:10.1101/2021.05.18.21257238
61. Stamm T, Bott N, Thwaites R, et al. Building a Value-Based Care Infrastructure in Europe: The Health Outcomes Observatory. *NEJM Catalyst Innovations in Care Delivery*. 2021;2(3)
62. Nilsson K, Bååthe F, Andersson AE, Wikström E, Sandoff M. Experiences from implementing value-based healthcare at a Swedish University Hospital – a longitudinal interview study. *BMC Health Serv Res*. 2017;17(1)doi:10.1186/s12913-017-2104-8
63. Erichsen Andersson A, Bååthe F, Wikström E, Nilsson K. Understanding value-based healthcare – an interview study with project team members at a Swedish university hospital. *Journal of Hospital Administration*. 2015;4(4):64. doi:10.5430/jha.v4n4p64
64. Steinmann G, Van De Bovenkamp H, De Bont A, Delnoij D. Redefining value: a discourse analysis on value-based health care. *BMC Health Serv Res*. 2020;20(1)doi:10.1186/s12913-020-05614-7
65. Steinmann G, Delnoij D, Van De Bovenkamp H, Groote R, Ahaus K. Expert consensus on moving towards a value-based healthcare system in the Netherlands: a Delphi study. *BMJ Open*. 2021;11(4):e043367. doi:10.1136/bmjopen-2020-043367
66. Delnoij DMJ, Steinmann G. Value-based care: requiring conceptual checks and international balances. *Eur J Public Health*. 2021;doi:10.1093/eurpub/ckab052
67. Dronkers EAC, Baatenburg De Jong RJ, Poel EF, Sewnaik A, Offerman MPJ. Keys to successful implementation of routine symptom monitoring in head and neck oncology with “ Healthcare Monitor ” and patients’ perspectives of quality of care. *Head Neck*. 2020;42(12):3590-3600. doi:10.1002/hed.26425

68. van Egdom LSE, Lagendijk M, van der Kemp MH, et al. Implementation of Value Based Breast Cancer Care. *Eur J Surg Oncol*. 2019/07/01/ 2019;45(7):1163-1170. doi:https://doi.org/10.1016/j.ejso.2019.01.007
69. Van Veghel D, Marteijn M, De Mol B. First results of a national initiative to enable quality improvement of cardiovascular care by transparently reporting on patient-relevant outcomes. *Eur J Cardiothorac Surg*. 2016;49(6):1660-1669. doi:10.1093/ejcts/ezw034
70. Van Veghel D, Daeter EJ, Bax M, et al. Organization of outcome-based quality improvement in Dutch heart centres. *European Heart Journal - Quality of Care and Clinical Outcomes*. 2019;doi:10.1093/ehjqcco/qcz021
71. Van Veghel D, Soliman-Hamad M, Schulz DN, Cost B, Simmers TA, Dekker LRC. Improving clinical outcomes and patient satisfaction among patients with coronary artery disease: an example of enhancing regional integration between a cardiac centre and a referring hospital. *BMC Health Serv Res*. 2020;20(1)doi:10.1186/s12913-020-05352-w
72. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*. 2019;6(1)doi:10.1186/s40537-019-0217-0
73. Madsen LB. *Data-driven healthcare: how analytics and BI are transforming the industry*. John Wiley & Sons; 2014.
74. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
75. Black N. Patient reported outcome measures could help transform healthcare. *BMJ*. 2013;346(jan28 1):f167-f167. doi:10.1136/bmj.f167
76. Gopal G, Suter-Crazzolara C, Toldo L, Eberhardt W. Digital transformation in healthcare - architectures of present and future information technologies. *Clin Chem Lab Med*. Feb 25 2019;57(3):328-335. doi:10.1515/cclm-2018-0658
77. Catalyst N. Healthcare big data and the promise of value-based care. *NEJM Catalyst*. 2018;4(1)
78. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020:m689. doi:10.1136/bmj.m689
79. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *The Lancet Digital Health*. 2021;3(4):e260-e265. doi:10.1016/s2589-7500(20)30317-4
80. Fridsma DB. Health informatics: a required skill for 21st century clinicians. *BMJ*. 2018;362:k3043. doi:10.1136/bmj.k3043
81. Dhindsa K, Bhandari M, Sonnadara RR. What's holding up the big data revolution in healthcare? *BMJ*. 2018;363:k5357. doi:10.1136/bmj.k5357
82. Kanzaki J, Tos M, Sanna M, Moffat DA, Monsell EM, Berliner KI. New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma. *Otol Neurotol*. Jul 2003;24(4):642-8; discussion 648-9. doi:10.1097/00129492-200307000-00019

Value based vestibular schwannoma care

Long-term quality of life of vestibular schwannoma patients: a longitudinal analysis

Olaf Neve
Jeroen Jansen
Radboud Koot
Mischa de Ridder
Peter Paul van Benthem
Anne Stiggelbout
Erik Hensen

Otolaryngol Head Neck Surg. 2023 Feb;168(2):210-217.
doi: 10.1177/01945998221088565

ABSTRACT

Objective

Vestibular schwannoma management aims to maintain optimal quality of life (QoL) while preventing severe sequelae of the tumor or its treatment. This study assessed long term QoL of vestibular schwannoma patients in relation to treatment modality and decisional regret.

Study Design

A longitudinal study, in which clinical and QoL data were used that were cross-sectionally acquired in 2014, and again in 2020 from the same patient group.

Setting

A tertiary expert center for vestibular schwannoma care in the Netherlands.

Methods

QoL was measured by the Penn Acoustic Quality of Life scale (PANQOL). Changes in time were assed using a linear mixed model. In addition the Decision Regret Scale was analyzed.

Results

Of 867 patients, 536 responded (62%), with a median follow-up of 11 years. All PANQOL subdomain scores remained stable over time and did not exceed minimal clinical important difference (MCID) levels. Time since treatment did not affect QoL. Patients had comparable average QoL scores and proportions of patients with changing QoL scores (i.e., exceeding the MCID) over time, irrespective of the received initial treatment. Female patients and those who required salvage therapy (either by radiotherapy or surgery) reported a lower QoL. The latter patient group reported the highest decisional regret.

Conclusion

On average, the long-term QoL of vestibular schwannoma patients is comparable for patients under active surveillance and those who have received active treatment, and remains stable over time. This suggests that, on average, preservation of QoL of vestibular schwannoma patients is feasible when adequately managed.

INTRODUCTION

Vestibular schwannomas (VSs) are rare and benign tumors arising in the cerebellopontine angle, typically causing hearing loss, tinnitus and balance disorders. In addition, facial numbness or pain, headache and facial paresis may occur.[1] A substantial minority of the tumors (22-48%) are progressive and may eventually lead to brainstem compression or increased intracranial pressure.[1] Management options comprise active surveillance, surgery or radiotherapy, and aim to prevent severe sequelae while maintaining patients' quality of life (QoL).[1] None of these modalities will improve symptoms, and all yield the risk of deterioration of hearing, balance and facial nerve function. Active surveillance is the management option for indolent tumors, whereas radiotherapy or surgical removal are indicated for progressive or large tumors, both resulting in >90% tumor control.[1, 2] Furthermore, the choice for a specific treatment option depends on additional tumor characteristics (e.g., localization and size) and patient-related factors such as the burden and type of symptoms and patient preference.[3]

Regardless of the treatment modality, a VS impacts a patient's QoL.[4-8] Ongoing dizziness and headache seem the most important determinants of poor QoL, as is large tumor size.[4, 9] QoL seems most affected directly after diagnosis.[4, 9] The effect of other determinants on general QoL, such as sex and education are not widely reported in VS research.

In 2010 Shaffer et al. developed the Penn Acoustic Neuroma Quality of life (PANQOL). This disease-specific questionnaire enabled more accurate QoL measurement than the generic QoL questionnaires.[7, 10-12] Several studies have published the results on the QoL of different treatment modalities.[4-8] Kerezoudis et al. have defined the minimal clinical important difference (MCID) of PANQOL scores, which advanced the interpretation of changes in PANQOL beyond the level of statistical significance.[13]

Little is known about the longitudinal impact of VS on QoL in the long term and whether patients regret the initial treatment decision. Longitudinal studies show changes within individuals over time, which cannot be detected in crossectional studies. This provides essential information since patients grow old with their tumor and the side effects of the chosen treatment. Previous studies on long-term QoL in VS patients were cross-sectional in design and showed no clinically relevant differences between treatment modalities.[5-7] Longitudinal studies lack long-term follow-up, do not use a disease-specific questionnaire or focus only on one treatment modality.[14-18] This study aims to find the long-term longitudinal QoL outcomes and evaluates decisional regret in VS patients.

METHODS

This longitudinal study was performed at the Leiden University Medical Center (LUMC), an expert center for VS in Leiden, the Netherlands. Data were first collected in 2014 for a cross-sectional study on QoL[4], and second between June and September 2020. The 'Medical Ethical Committee Leiden, Den Haag, Delft' waived the necessity for medical ethical approval under Dutch law and approved the study regarding data handling and privacy regulations (N19.112).

Patients who had participated in 2014 were reapproached for participation by mail or email. Inclusion criteria were age ≥ 18 years, and a unilateral VS. Patients with other skull base pathologies were excluded. All patients were referred to the LUMC between 2004 and 2014.

Patients were asked to complete the PANQOL questionnaire. This validated questionnaire measures VS-related QoL and consists of 26 items divided over seven subdomains (hearing, balance, face, pain, energy, anxiety, and general health).[10, 11] The Likert-scale answers were summed per subdomain and recoded to range from 0 to 100, with higher scores indicating better QoL. A total score was calculated as the arithmetic mean of the subdomain scores. Incomplete answers were excluded on a subdomain level. In addition, patients completed the Decision Regret Scale(DRS), a four-item validated questionnaire.[19] A total score was calculated from 0 to 100, with higher scores indicating high regret. Scores of 0 were defined as no regret and >50 as considerable regret. Incomplete questionnaires were excluded from the analysis. Furthermore, patients completed a short questionnaire with demographic parameters about sex, age, occupation and education level. Statistics Netherlands (CBS) definition for low, middle and high education level was used, which follows the international standard classification of education.[20]

Tumor size, treatment modality, time since diagnosis and treatment were retrospectively acquired from patient records. Treatment modality was categorized as active surveillance, surgery, radiotherapy, or both surgery and radiotherapy. Patients only received both surgery and radiotherapy if the initial treatment did not result in adequate tumor control: patients in whom radiotherapy failed underwent salvage surgery, and patients in whom surgery failed underwent salvage radiotherapy. Tumor size was scored at the start of treatment or in the case of active surveillance at the time of the first questionnaire in 2014 using the reporting system proposed by Kanzaki et al.[21] Categories large and giant were merged because of the small number of patients in these categories
Statistical analyses were performed in R version 4.0.5 using Rstudio 1.3.959 (Rstudio, PBC, Boston) and data plotted using the package *ggplot2*. Means and standard devia-

tions were calculated for normally distributed variables and medians and interquartile ranges(IQR) for non-normally distributed continuous variables. For categorical variables, counts and frequencies were calculated. A non-responder analysis was performed to check for differences in baseline characteristics using unpaired t-tests.

Differences in PANQOL subdomains between 2014 and 2020 were tested per treatment modality with a paired T-test and Bonferroni correction for multiple testing to prevent type I error. In addition, the differences were compared to the median anchored subdomain specific minimal clinically important differences(MCID) reported by Kerezoudis et al.[13] A difference smaller than the MCID was defined as stable and larger differences as either deterioration or improvement of QoL. Differences in decisional regret between treatment modalities were tested pairwise using the Wilcoxon rank test with Bonferroni correction.

Long-term effects of time since treatment and treatment modality on QoL were analyzed using a linear mixed model (R-package *nlme*) to account for repeated measurement data. Model assumptions were visually checked. The total PANQOL score was the dependent variable, with two measurements per patient (2014 and 2020). Covariates such as age, education level, sex, tumor size were step-wise included in the model and model selection was based on Akaike information criterion, as were interactions between time since treatment and other covariates. In the final model, random intercepts were used with fixed variables. All tests were two-sided and p-values <0.05 were considered statistically significant. Patients treated (either by surgery, radiotherapy or both) after 2014, i.e., between the two measurements, were analyzed separately. Because of the small sample size, no further statistical analysis was performed on these data.

RESULTS

In 2014, 913 patients completed one or more PANQOL subdomains.[4] In 2020, 867/913 patients were still alive and were reapproached, of whom 536 responded (62% response rate), as shown in Figure 1. After the first measurement in 2014, 36 patients have been actively treated with either surgery or radiotherapy. These patients were analyzed separately. In total, 487 patients completed one or more subdomains of the PANQOL. Figure 1 shows the number of patients who completed a specific PANQOL subdomain in 2014 and 2020.

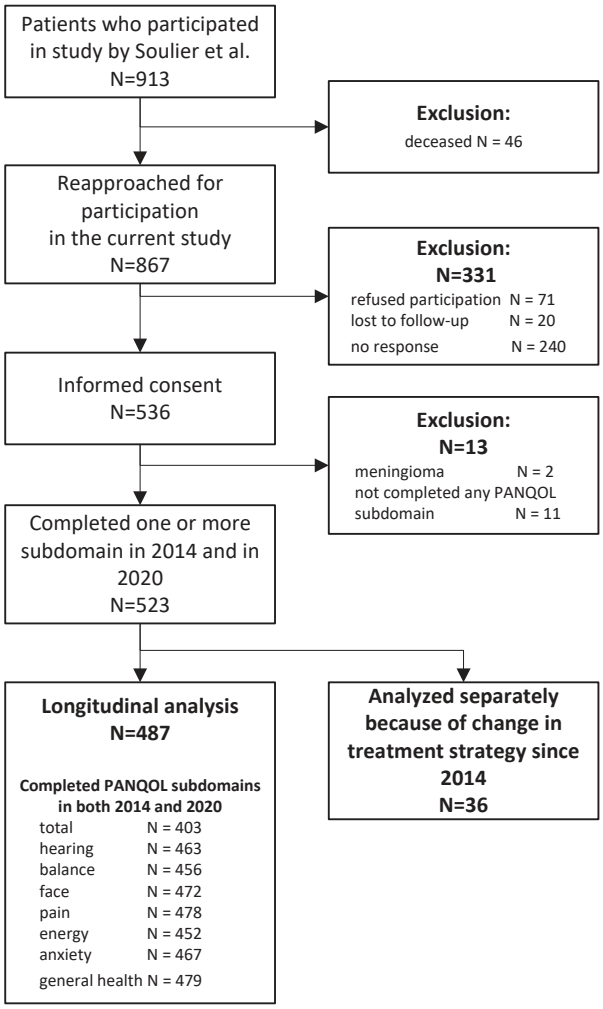


Figure 1. Flowchart of study participants. The number of patients who completed the Penn Acoustic Neuroma Quality of Life (PANQOL) questionnaire both in 2014 and 2020 are shown in the last box per subdomain.

In the non-responder group (N=331), 240 (62%) did not respond for unknown reasons, 20 (6%) were lost to follow-up, and 71 (24%) declined participation. The most frequent reason for declining participation was lack of time (31%), followed by health problems other than VS (14%), 28% did not provide a reason. Responders were on average nine months younger, had higher education levels (high-level education 21% vs. 14%) and had received surgery more often (22% vs. 10%) than non-responders.

The median time since treatment was 10 years, and since diagnosis 11 years. Patients who underwent surgery were on average younger, more often female, and had larger tumors at the start of the treatment (Table 1).

Table 1. Baseline characteristics 2020

		Treatment modality			
	Total	Active surveillance	Surgery	Radiotherapy	Surgery + Radiotherapy
N	487	246	179	47	15
Age (sd)	67.4 (10.8)	69.6 (10.6)	64.1 (10.1)	69.9 (10.7)	62.9 (13.8)
Women (%)	226 (46.3)	102 (41.3)	99 (55.3)	18 (38.3)	7 (46.7)
Education (%)					
low	157 (32.2)	93 (37.8)	52 (29.1)	8 (17.0)	4 (26.7)
middle	148 (30.4)	66 (26.8)	60 (33.5)	15 (31.9)	7 (46.7)
high	182 (37.4)	87 (35.4)	67 (37.4)	24 (51.1)	4 (26.7)
Time since					
treatment median (range)	10 (7-21)	-	11 (7-17)	9 (7-16)	9 (7-16)
diagnosis median (range)	11 (7-21)	10 (7-21)	12 (7-21)	10 (8-18)	11 (7-16)
Kanzaki at treatment* (%)					
intrameatal	135 (28)	117 (48)	14 (8)	4 (9)	0
small (0-10mm)	115 (24)	69 (28)	35 (20)	11 (23)	0
medium (11-20mm)	140 (29)	53 (22)	58 (32)	22 (47)	7 (47)
moderately large (21-30mm)	62 (13)	5 (2)	41 (23)	10 (21)	6 (40)
large (31-40 mm)	25 (5)	1 (<1)	23 (13)	0	1 (7)
giant (>40mm)	9 (2)	0	8 (4)	0	1 (7)
Missing	1 (<1)	1 (<1)	0	0	0
Decision regret scale					
median (IQR)	10 (0-25)	0 (0-20)	15 (5-25)	10 (0-30)	25 (18-35)

* for active surveillance tumor size in 2014 is used. IQR = Interquartile range, sd = standard deviation

The paired PANQOL scores of 2014 and 2020 are shown in Figure 2. Only balance scores showed deterioration over time in both the active surveillance (-6.3 95% confidence interval (CI) -3.9, -9.3) and surgery groups (-4.8 95%CI -1.9, 7.7). Both differences did not exceed the MCID of 14 points. The group receiving both surgery and radiotherapy showed a non-significant trend of deteriorating scores at all subdomains.

When the changes over time in PANQOL scores were compared with the MCID (12.5 points), the majority of all patients (n= 278, 69%) was stable, i.e., showed a difference of less than 12,5 points. In the active surveillance group, 36 patients (17.9%) had a deterioration in the overall PANQOL score, whereas 27 (13.4%) patients reported an improvement in PANQOL. The majority (135 patients; 67.2%) had a stable PANQOL

score. In the surgery group, 23 patients (15.7%) deteriorated, 18 (11.9%) improved and 108 (71.5%) were stable. In the radiotherapy group 4 patients (11.1%) deteriorated, 4 (11.1%) improved and 27 (75.0%) remained stable. These differences between treatment strategy groups are small and not statistically significant (χ^2, p=0.8). In addition, we found no baseline differences between patients with a stable long term QoL and those with decreasing or increasing QoL scores over time.

The effects of sociodemographic and clinical characteristics on the total PANQOL score were assessed using a linear mixed model (Table 2. and Figure 3). There was no statistically significant association between time since treatment and total PANQOL score, and associations between other covariates and PANQOL score were not dependent on the time since treatment. Therefore, the interaction terms were omitted from the final model.

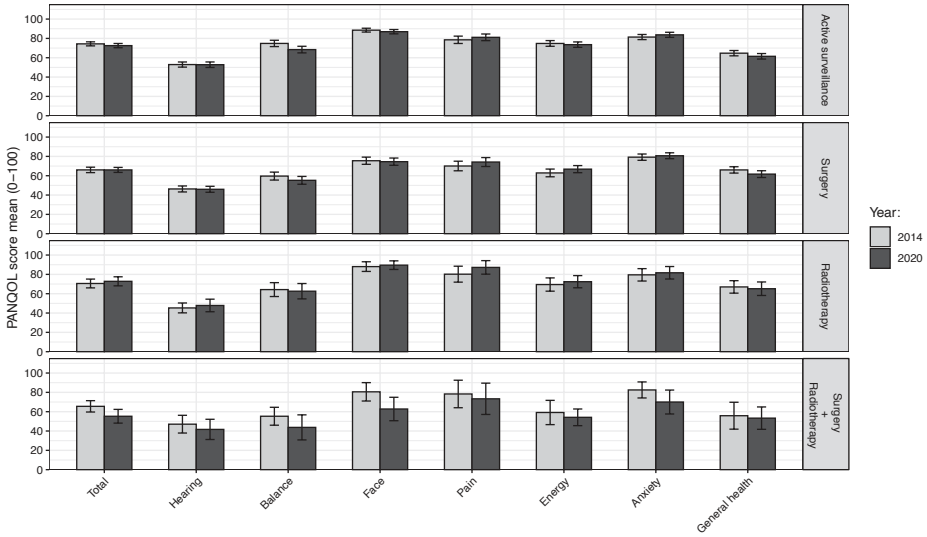


Figure 2. Paired unadjusted mean Penn Acoustic Neuroma Quality of Life (PANQOL) scores 2014 and 2020 per treatment modality. Error bars indicate 95% CIs of the means. Higher scores indicate better quality of life. The total score is the arithmetic mean of the subdomains.

A high level of education was significantly related to a better QoL than a low level of education (7.2 point difference on the PANQOL, 95%CI 3.4, 10.6), although the difference did not reach the MCID level. Also, women had a significantly lower QoL than men (-6.0, 95%CI -8.9, -3.1). This sex-related difference occurred across all treatment modalities. The lower total scores of women for active surveillance (-6.1 95%CI -10.8, -1.3) and surgery (-9.3 95%CI -14.6,-4.1) were smaller than the MCID of 12.5. For radiotherapy (-12.0 95%CI -23.3, -0.6) and radiotherapy + surgery (-14.8 95%CI -29.3, -0.3), the differences

Table 2. Linear mixed model

	Dependent variable: PANQOL total score		
	(1) univariate	**(2) multivariate**	
	Estimated means (95% CI)	*Estimated means (95%CI)*	*Marginal means (95%CI)*
Time since treatment			
0-1 years	70.0 (66.4, 73.7)	66.5 (62.4, 70.7)	*reference*
2-4 years	71.4 (69.2, 73.7)	68.4 (65.4, 71.4)	1.8 (-3.5, 7.2)
5-7 years	69.2 (67.0, 71.4)	65.8 (62.8, 68.9)	-0.7 (-5.1, 3.7)
≥ 8 years	69.8 (68.1, 71.5)	66.9 (64.2, 69.9)	0.4 (-4.5, 5.2)
Sex			
men	73.7 (71.8, 75.6)	69.9 (66.9, 72.9)	*reference*
women	65.7 (63.4, 68.0)	63.9 (60.8, 67.0)	**-6.0 (-8.9, -3.1)**
Education			
low	67.1 (64.5,69.7)	64.1 (60.6, 67.7)	*reference*
middle	68.1 (65.5,70.8)	65.4 (62.1, 68.8)	1.3 (-3.1, 5.7)
high	74.5 (72.1, 76.9)	71.2 (67.9, 74.4)	**7.0 (2.7, 11.3)**
Treatment modality			
active surveillance	73.4 (71.3, 75.5)	73.3 (70.5, 76.1)	*reference*
surgery	65.7 (63.3, 68.2)	65.8 (63.2, 68.4)	**-7.5 (-12.4, -2.6)**
radiotherapy	71.9 (67.1, 76.6)	68.8 (63.8, 73.7)	-4.5 (-11.5, 2.4)
surgery + radiotherapy	61.0 (52.6, 69.3)	59.8 (51.7, 68.0)	**-13.4 (-24.8, -2.1)**
Kanzaki at treatment			
intrameatal	71.7 (68.9, 74.6)	65.7 (61.8, 69.6)	*reference*
small (0-10mm)	69.4 (66.4, 72.4)	65.3 (61.6, 69.0)	-0.4 (-5.8, 5.1)
medium (11-20mm)	70.1 (67.5,72.8)	68.2 (65.0, 71.3)	2.5 (-3.1, 8.0)
moderately large (21-30mm)	69.9 (66.0, 73.8)	70.4 (66.3, 74.6)	4.8 (-2.5, 12.0)
large + giant (>30 mm)	64.0 (58.5, 69.6)	65.0 (59.1, 70.9)	-0.7 (-9.9, 8.5)
Observations	-	868	
Log-Likelihood	-	-3,5249	
Akaike Inf. Crit.	-	7,132	

Linear mixed model with total PANQOL total score as dependent variable.
Model 1 shows the estimated means of univariate analyses for every variable, with random intercepts. Model 2 shows estimated means of multivariate analysis in which all variables were included, with random intercepts. In the right column the marginal means are shown compared to the reference category. Significant differences are shown in bold. In both models age in 2020 was a covariate. The 95% confidence intervals are shown between the brackets ().

were close to or above the MCID. Although individual sociodemographic characteristics did not result in PANQOL scores exceeding the MCID level, a combination of different sociodemographic factors may. For example, males with a high level of education tended to have a higher PANQOL total score (+13.2) than females with a low level of education.

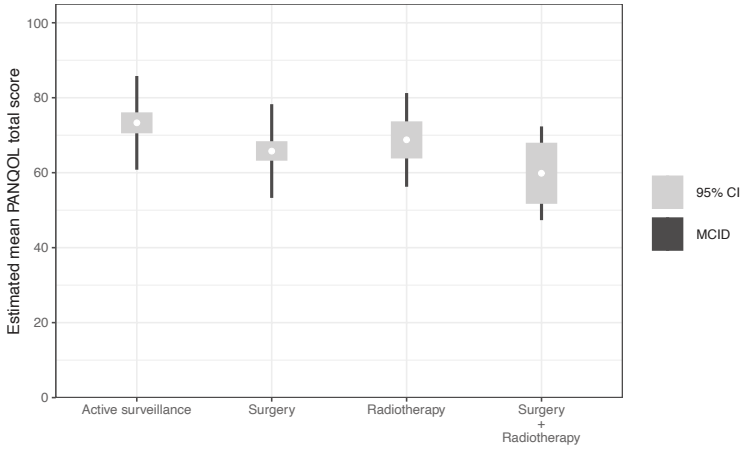


Figure 3. Estimated means of the Penn Acoustic Neuroma Quality of Life (PANQOL). Results of the linear mixed model of the total PANQOL score per treatment modality corrected for confounding factors (age, sex, education level, time since treatment, tumor size at treatment) per treatment.

In addition, receiving a disability pension (N=37) was associated with lower PANQOL total scores. The mean difference in PANQOL total scores of full-time employed patients (N=205) and patients with a disability pension were 25.3(95%CI 20.5, 30.1), exceeding the 12.5 MCID. Furthermore, differences of PANQOL total scores of unemployed (N= 7; -13.0,95%CI 1.6, -27.6) and voluntary unemployed (N=36;-13.3,95%CI -8.0, -18.7) exceeded the MCID.

Total PANQOL scores in the surgery (-7.5 95%CI -11.2, -3.8) and surgery+radiotherapy (-13.5, 95%CI -22.1, -4.8) groups were lower than active surveillance group. The difference with the group receiving both surgery and radiotherapy exceeded the 12.5-point MCID. When analyzing the separate subdomains, differences were found for balance after surgery (-14.8) and radiotherapy (-15.4), exceeding the MCID of 14. Differences in anxiety in the radiotherapy group (-18) exceeded the MCID (13).

Decision regret was analyzed per treatment modality (Table 1.). In the active surveillance group, 52% scored 0, indicating no regret at all, and 2% scored >50, indicating considerable regret. After surgery, the median score was 15 (IQR 5-25), while 23% had no regret and 7% had considerable regret. After radiotherapy, the median score was 5 (IQR 0-20), 49% had no regret and 6% considerable regret. Patients in the surgery + radiotherapy group had the highest DRS scores, with a median of 25 (IQR 18-35). Only 7% had no regret in this patient group, while 20% had considerable regret. Compared to active surveillance, surgery ($p<0.0001$) and surgery + radiotherapy ($p=0.002$) scored significantly worse. The difference between surgery and radiotherapy was not statistically significant.

Between the two measurements, 36 patients were actively treated, of whom 28 completed all PANQOL questions in 2014 and 2020. Of the patients receiving surgery (N=9), the QoL of one patient deteriorated (11%), the other patients remained stable (i.e., within the MCID limits) over time. In the radiotherapy group, 7(36.8%) deteriorated, 3(15.8%) improved and 9(47.4%) remained stable. Median DRS in the groups were 10 and 20 for surgery and radiotherapy, respectively.

DISCUSSION

This longitudinal study showed that although the individual variation is considerable, on average, the QoL of VS patients remains stable over time and is comparable for all treatment modalities, except for patients requiring salvage therapy after initial therapy failure. These patients seemed to have a lower QoL that declined over time. This group also has the highest decision regret, while in other groups, decision regret is low.

Although the disease-specific QoL remained stable on average, a minority of the VS patients reported changing PANQOL scores over time. There were no significant differences in the proportion of patients experiencing decreased or increased QoL between treatment modalities. In addition, we could not find reliable predictors for either improvement or deterioration of QoL over time.

Since the development of the disease-specific PANQOL, several large cross-sectional studies have been performed assessing QoL in VS patients.[4-8, 22] None have shown clinically relevant differences between treatment modalities. In agreement with the current study, McLaughlin et al.[22] and Carlson et al.[5] reported a slightly lower QoL after surgery that did not exceed the MCID. Previous short-term longitudinal studies that used the PANQOL showed no differences in QoL outcomes per treatment modality and no differences over time, as was observed in this study.[14, 16]

A clinically relevant lower QoL was found in patients requiring multiple treatments (i.e. salvage therapy by radiotherapy or surgery after initial therapy failure). Although the group size was limited, the differences were statistically significant. In addition, this group seemed to have a declining trend over time in all PANQOL subdomains. This finding is in agreement with the study by Carlson et al., who reported a statistically significant difference between active surveillance and multimodality treatment.[6]

Importantly, as in previous studies, treatment groups were not similar at baseline in this study. For example, patients undergoing surgery tended to be younger and had larger tumors, pa-

tients requiring both radiotherapy and surgery had failed initial treatment. These differences reflect the indications for specific treatment modalities at our center. Although preservation of QoL is an important goal of VS management, other factors (such as tumor progression or size) usually determine the necessity for active intervention. The choice of treatment modality is also not determined by its intrinsic contribution to a patient's QoL. Moreover, the finding that long-term QoL is comparable for all three management strategies (radiotherapy, surgery, and active surveillance), does not mean that treatments are interchangeable from a QoL perspective or that the choice or timing of treatment is of little relevance. Rather, the comparability of long-term quality life after different VS treatment strategies in retrospect can be viewed as the result of personalized treatment decisions, deploying a specific therapy in a specific patient at a specific moment in the course of the disease.

The current study shows that even >10 years after treatment, QoL is stable across modalities, which supports the results of two cross-sectional studies on long-term QoL.[5, 8] Patients in the active surveillance and surgery group had a minor deterioration of the balance subdomain, however not exceeding the MCID. This minor decline was not observed in two short-term longitudinal studies.[14, 16] The deterioration could be due to an aging effect, which might cause increased balance problems combined with the VS. Other longitudinal studies have reported contradicting changes in anxiety, especially in patients undergoing surgery.[14, 16] In the current study, no changes in anxiety were observed. It might be possible that anxiety is affected shortly after diagnosis and/or treatment and remains stable over time afterward.

The current study identified several factors associated with worse long-term QoL in VS patients besides the requirement of salvage treatment: female sex and disability pension. Sex-related differences in QoL were observed, with lower QoL in women specifically for the balance and anxiety subdomains. To our knowledge, the difference in QoL between male and female patients has not been described before in VS. The PANQOL validation study identified no sex-related differences.[11] Many studies have corrected for sex but did not report the effect of sex on the QoL.[4, 6-8, 14, 16] However, this sex difference (women reporting a lower QoL) has been reported in other diseases and in the general population too, which can only be partly explained by differences in social-economic status.[23, 24,25] Other possible explanations might be sex-related differences in reporting symptoms or an actual difference in the disease-related QoL.[26]

Decision regret was low in the active surveillance group and slightly higher in radiotherapy and surgery groups. One previous study in VS patients supported the findings in this study, albeit using non-validated questions, reporting 97%, 96% and 85% satisfaction for active surveillance, radiotherapy, and surgery, respectively.[27] A systematic

review in various diseases (mainly oncology) showed a mean DRS score of 16.5.[28] In the current study, only patients requiring salvage therapy, and thus receiving both surgery and radiotherapy during the course of the disease, had a higher decisional regret. This is not surprising as in these patients initial treatment had failed.

This study has some limitations. The retrospective design carries an inherent risk of selection bias, although this design was inevitable for gathering long-term longitudinal results of patients diagnosed before the development of the disease-specific questionnaire. The participating patients were diagnosed or/and treated in one center, and it might be possible that a selection of patients suffering from the sequelae was more likely to participate. Furthermore, the group of non-responders might introduce selection bias since this group had different demographic characteristics.[29] In addition, the patient group requiring salvage therapy was small because recurrences are relatively rare.

CONCLUSION

This longitudinal study shows that QoL in VS patients is stable over time and that different management strategies (surgery, radiotherapy and active surveillance) result in comparable long-term QoL outcomes. There is, however, considerable individual variation. Factors associated with a decreased long-term QoL in VS patients are female sex, receiving a disability pension, and the need for salvage treatment after initial therapy failure.

Acknowledgments

We thank professor Hein Putter of the LUMC statistics department for his advice and help in choosing the right statistical approach.

REFERENCES

1. Carlson ML, Link MJ. Vestibular Schwannomas. *N Engl J Med*. 2021;384(14):1335-1348. doi:10.1056/nejmra2020394
2. Møller MN, Hansen S, Miyazaki H, Stangerup S, Caye-Thomasen P. Active Treatment is Not Indicated in the Majority of Patients Diagnosed with a Vestibular Schwannoma : A Review on the Natural History of Hearing and Tumor Growth. *Current Otorhinolaryngology Reports*. 2014;2:242-247. doi:10.1007/s40136-014-0064-7
3. Neve OM, Soulier G, Hendriksma M, et al. Patient-reported factors that influence the vestibular schwannoma treatment decision: a qualitative study. *Eur Arch Otorhinolaryngol*. 2020;doi:10.1007/s00405-020-06401-0
4. Soulier G, Van Leeuwen BM, Putter H, et al. Quality of Life in 807 Patients with Vestibular Schwannoma: Comparing Treatment Modalities. *Otolaryngology–Head and Neck Surgery*. 2017;157(1):92-98. doi:10.1177/0194599817695800
5. Carlson ML, Tveiten OV, Driscoll CL, et al. Long-term quality of life in patients with vestibular schwannoma: an international multicenter cross-sectional study comparing microsurgery, stereotactic radiosurgery, observation, and nontumor controls. *J Neurosurg*. 2015;122:833-842. doi:10.3171/2014.11.JNS14594
6. Carlson ML, Tombers NM, Kerezoudis P, Celda MP, Lohse CM, Link MJ. Quality of Life Within the First 6 Months of Vestibular Schwannoma Diagnosis With Implications for Patient Counseling. *Otol Neurotol*. 2018;39(10):e1129-e1136. doi:10.1097/mao.0000000000001999
7. Lodder WL, Van Der Laan BFAM, Lesser TH, Leong SC. The impact of acoustic neuroma on long-term quality-of-life outcomes in the United Kingdom. *Eur Arch Otorhinolaryngol*. 2018-03-01 2018;275(3):709-717. doi:10.1007/s00405-018-4864-0
8. Robinett ZN, Walz PC, Miles-Markley B, Moberly AC, Welling DB. Comparison of Long-term Quality-of-Life Outcomes in Vestibular Schwannoma Patients. *Otolaryngology–Head and Neck Surgery*. 2014;150(6):1024-1032. doi:10.1177/0194599814524531
9. Carlson ML, Tveiten ØV, Driscoll CL, et al. What drives quality of life in patients with sporadic vestibular schwannoma? *The Laryngoscope*. 2015;125(7):1697-1702. doi:10.1002/lary.25110
10. van Leeuwen BM, Herruer JM, Putter H, Jansen JC, van der Mey AG, Kaptein AA. Validating the Penn Acoustic Neuroma Quality Of Life Scale in a sample of Dutch patients recently diagnosed with vestibular schwannoma. *Otol Neurotol*. Jul 2013;34(5):952-7. doi:10.1097/MAO.0b013e31828bb2bb
11. Shaffer BT, Cohen MS, Bigelow DC, Ruckenstein MJ. Validation of a disease-specific quality-of-life instrument for acoustic neuroma. *The Laryngoscope*. 2010;120(8):1646-1654. doi:10.1002/lary.20988
12. Kristin J, Glaas MF, Schipper J, et al. Patient quality of life after vestibular schwannoma removal: possibilities and limits to measuring different domains of patients' wellbeing. *Eur Arch Otorhinolaryngol*. 2019;276(9):2441-2447. doi:10.1007/s00405-019-05499-1
13. Kerezoudis P, Yost KJ, Tombers NM, Celda MP, Carlson ML, Link MJ. Defining the Minimal Clinically Important Difference for Patients With Vestibular Schwannoma: Are all Quality-of-Life Scores Significant? *Neurosurgery*. 2019;85(6):779-785. doi:10.1093/neuros/nyy467
14. Miller LE, Brant JA, Naples JG, Bigelow DC, Lee JYK, Ruckenstein MJ. Quality of Life in Vestibular Schwannoma Patients: A Longitudinal Study. *Otol Neurotol*. 2020-02-01 2020;41(2):e256-e261. doi:10.1097/mao.0000000000002445

15. Windisch P, Tonn J-C, Fürweger C, et al. Longitudinal Changes of Quality of Life and Hearing Following Radiosurgery for Vestibular Schwannoma. *Cancers (Basel)*. 2021;13(6):1315. doi:10.3390/cancers13061315
16. Carlson ML, Barnes JH, Nassiri A, et al. Prospective Study of Disease-Specific Quality-of-Life in Sporadic Vestibular Schwannoma Comparing Observation, Radiosurgery, and Microsurgery. *Otol Neurotol*. 2021;42(2):e199-e208. doi:10.1097/mao.0000000000002863
17. Breivik CN, Varughese JK, Wentzel-Larsen T, Vassbotn F, Lund-Johansen M. Conservative Management of Vestibular Schwannoma—A Prospective Cohort Study: Treatment, Symptoms, and Quality of Life. *Neurosurgery*. 2012-05-01 2012;70(5):1072-1080. doi:10.1227/neu.0b013e31823f5afa
18. Park SS, Grills IS, Bojrab D, et al. Longitudinal assessment of quality of life and audiometric test outcomes in vestibular schwannoma patients treated with gamma knife surgery. *Otol Neurotol*. Jun 2011;32(4):676-9. doi:10.1097/MAO.0b013e3182138fc5
19. Brehaut JC, O'Connor AM, Wood TJ, et al. Validation of a Decision Regret Scale. *Med Decis Making*. 2003;23(4):281-292. doi:10.1177/0272989x03256005
20. Statistics Netherlands (2017) Standaard onderwijsindeling 2016. Den Haag. https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/onderwijs-en-beroepen/standaard-onderwijsindeling--soi--/standaard-onderwijsindeling-2016 Accessed 8 Mar 2021
21. Kanzaki J, Tos M, Sanna M, Moffat DA, Monsell EM, Berliner KI. New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma. *Otol Neurotol*. Jul 2003;24(4):642-8; discussion 648-9. doi:10.1097/00129492-200307000-00019
22. McLaughlin EJ, Bigelow DC, Lee JY, Ruckenstein MJ. Quality of life in acoustic neuroma patients. *Otol Neurotol*. Apr 2015;36(4):653-6. doi:10.1097/mao.0000000000000674
23. Reeves MJ, Bushnell CD, Howard G, et al. Sex differences in stroke: epidemiology, clinical presentation, medical care, and outcomes. *The Lancet Neurology*. 2008;7(10):915-926. doi:10.1016/s1474-4422(08)70193-5
24. Pettersen KI, Reikvam A, Rollag A, Stavem K. Understanding sex differences in health-related quality of life following myocardial infarction. *Int J Cardiol*. 2008;130(3):449-456. doi:10.1016/j.ijcard.2007.10.016
25. Cherepanov D, Palta M, Fryback DG, Robert SA. Gender differences in health-related quality-of-life are partly explained by sociodemographic and socioeconomic variation between adult men and women in the US: evidence from four US nationally representative data sets. *Qual Life Res*. 2010;19(8):1115-1124. doi:10.1007/s11136-010-9673-x
26. Gijsbers van Wijk CMT, Kolk AM. Sex differences in physical symptoms: The contribution of symptom perception theory. *Soc Sci Med*. 1997;45(2):231-246. doi:10.1016/s0277-9536(96)00340-1
27. Carlson ML, Tveiten ØV, Lund-Johansen M, Tombers NM, Lohse CM, Link MJ. Patient Motivation and Long-Term Satisfaction with Treatment Choice in Vestibular Schwannoma. *World Neurosurg*. 2018-06-01 2018;114:e1245-e1252. doi:10.1016/j.wneu.2018.03.182
28. Becerra Pérez MM, Menear M, Brehaut JC, Légaré F. Extent and Predictors of Decision Regret about Health Care Decisions. *Med Decis Making*. 2016;36(6):777-790. doi:10.1177/0272989x16636113
29. Johnson TP. Response Rates and Nonresponse Errors in Surveys. *JAMA*. 2012;307(17):1805. doi:10.1001/jama.2012.3532

The impact of vestibular schwannoma and its management on employment

Olaf Neve
Jeroen Jansen
Andel van der Mey
Radboud Koot
Mischa de Ridder
Peter Paul van Benthem
Anne Stiggelbout
Erik Hensen

Eur Arch Otorhinolaryngol. 2022 Jun;279(6):2819-2826.
doi: 10.1007/s00405-021-06977-1

ABSTRACT

Background

Employment is an important factor in quality of life. For vestibular schwannoma (VS) patients employment is not self-evident, because of the sequelae of the disease or its treatment and their effects on daily life.

Objectives

This study assessed employment status, sick leave (absenteeism) and being less productive at work (presenteeism) in the long-term follow-up of VS patients, and evaluated the impact of treatment strategy (active surveillance, surgery or radiotherapy).

Methods

A cross-sectional survey study was performed in a tertiary university hospital in the Netherlands. Patients completed the iMTA-post productivity questionnaire (iPCQ). Employment status was compared to that of the general Dutch population. Employment, absenteeism and presenteeism were compared between patients under active surveillance, patients after radiotherapy and post-surgical patients.

Result

In total 239 patients participated, of which 67% were employed at the time of the study. Only 13% had a disability pension, which was comparable to the age-matched general Dutch population. The proportion of patients with absenteeism was 8%, resulting in a 4% reduction of working hours. Presenteeism was reported by 14% of patients, resulting in a 2% reduction of working hours. The median number of working hours per week was 36, and since the diagnosis, these hours had been reduced by 6%. There were no significant differences between treatment modalities.

Conclusion

On average, long-term employment status and working hours of VS patients are comparable to the age-matched general population. Treatment strategies do not seem to differentially impact on long-term employment of VS patients.

INTRODUCTION

Despite the benign character of vestibular schwannomas (VS), the tumor can have a substantial impact on patients' lives. Typically the tumor causes hearing loss, tinnitus and balance disorders. However, headache, facial numbness, facial pain or facial paresis may also occur.[1] These symptoms can reduce health-related quality of life (HRQL), and VS patients have reported even worse HRQL than patients with chronic diseases or head and neck cancer.[2]

Multiple treatment options exist for sporadic VS, which can be subdivided in three broad categories: active surveillance, radiotherapy and surgery. The treatment of choice depends on tumor characteristics (i.e. tumor size, progression), patient characteristics (i.e. age, symptoms, patient preference) and probably also on other factors such as availability in the particular hospital. The aim of all three strategies is tumor control, but the way this is achieved differs: with surgery the tumor mass is removed, either totally, near totally or subtotally.[3] Radiotherapy is aimed at arrest of tumor progression, while the tumor remains in situ. Active surveillance does not intervene with the tumor, but relies on the observation that a small majority of VSs does not show progression once detected.[3] In this strategy, active therapy is reserved for progressive disease. Importantly, none of these treatment options is well suited for curing hearing or vestibular function loss, and all confer a risk to hearing, balance, trigeminal or facial nerve function. As VS in general is not a life threatening disease when adequately managed, most patients will live with their tumor, the associated symptoms and/or the sequelae of therapy for prolonged periods of time. As such, a VS can be viewed as a chronic disease and the impact on HRQL may thus be lifelong. Because of this, HRQL is one of the guiding principles in VS management.

An important aspect of HRQL that is often overlooked is a patients' ability to participate in professional life.[4] In VS patients professional performance is not self-evident and may be impacted by associated symptoms and the reported social restrictions that patients experience because of them.[5,6] A patients' ability to acquire or maintain a job is critical from both a societal and personal perspective. Being employed is not only of paramount importance from an economic perspective, but also associated with increased self-esteem and self-worth.[7] Furthermore, work provides relationships, social connections and a higher level of social status, and having secure employment helps people to prevent developing illness.[8-10] Conversely, being unemployed is associated with a poorer physical and mental health.[9]

In previous studies, VS patients' employment rates varied between 45 and 80%.[11-14] Most studies have assessed the employment rate pre- and post-surgery or radiotherapy, and only one has assessed employment rates in VS patients under active surveillance. In addition to employment rates, the reduced productivity while having a job (presenteeism) and sick leave from a job (absenteeism) are important additional aspects of professional participation that have not yet been investigated in VS.

This study assesses the long-term effects of VS on employment and productivity and analyzes the long-term impact on employment of surgery, radiotherapy and active surveillance. Furthermore, determinants of unemployment and reduced productivity are evaluated.

METHODS

This cross-sectional study was part of a more extensive study on long-term VS outcomes in the Netherlands. Participating patients from a questionnaire study in 2014 were reapproached for participation.[15] All patients were diagnosed with unilateral VS between 2003 and 2014. Patients were diagnosed and/or treated at Leiden University Medical Center, an expert center for VS offering different management options, including active surveillance, surgery (mostly through translabyrinthine or retrosigmoidal approach) and fractionated stereotactic radiotherapy. Patients who prefer radiosurgery were referred to another clinic.

For this study, all patients between 18 and 67 years were included, since the retirement age is currently elevated stepwise from 65 to 67 years between 2013 and 2024 in the Netherlands. Exclusion criteria were other skull base pathologies and insufficient proficiency in the Dutch language to complete the questionnaires. After providing informed consent, patients could complete questionnaires electronically or on paper between April and September 2020.

Questionnaires

Employment status was assessed using the Productivity Costs Questionnaire (iPCQ). The iPCQ measures productivity loss from the societal perspective and contains three different modules: absenteeism, presenteeism, and productivity loss in unpaid/volunteer work. The first two modules are validated and the validation of the productivity loss for unpaid work is still in progress.[16] Productivity losses were calculated in hours following the manual.[16] The recall period used in the questionnaire is four weeks.[16] In

addition, questions about productivity before diagnosis were asked, while recognizing the reduced reliability due to the prolonged recall period.

Sex, age, time since the treatment and tumor size at diagnosis were acquired from the electronic patient records. Tumor size was scored at diagnosis using the reporting system proposed by Kanzaki et al.[17] The definition of Statistics Netherlands (CBS) for low, middle and high education level was used, which follows the international standard classification of education.[18] Frequencies were calculated for categorical variables and means and standard deviation (sd) for normally distributed numerical variables and medians and interquartile ranges (IRQ) for not-normally distributed numerical variables. Baseline characteristics of responders and non-responders were checked in a non-responder analysis.

Employment status of patients aged 45-65 years was compared to the general Dutch population aged between 45-65 years using a chi-squared test. This age category was the best matched age group to the study population available in the public data of Statistics Netherlands.[19] Employment status per treatment modality was compared using logistic regression. Sex, age, and educational level were included in the regression to correct for potential confounding. The goodness of fit was checked with a model chi-squared test.

Absenteeism and productivity in hours were compared to the general Dutch population in the third quarter of 2020. The effect of treatment modalities on productivity in hours per week was assessed in a linear regression. Sex, age and educational level were included to correct for potential confounding. Model assumptions were visually checked. The differences pre-and post-diagnosis of productivity in hours per week were analyzed per treatment modality.

All analyses were performed using SPSS version 26 (IBM SPSS Inc., Armonk, USA). A p-value <0.05 was considered statistically significant. Incomplete questionnaire modules (absenteeism, presenteeism, unpaid work) were excluded from the analysis. According to our power calculation, a minimum number of 37 participants per group were needed to detect a difference in the productivity of eight hours in four weeks (α=0.05, 1-β=0.8, assuming σ=6).

RESULTS

In total, 402 patients were approached for participation, of whom 243 (60%) provided informed consent. There were no significant differences between responders and non-

responders regarding age, sex and educational level. In the responder group, two patients were excluded because histology showed meningioma rather than schwannoma and two patients did not complete any questions after providing informed consent (Figure 1).

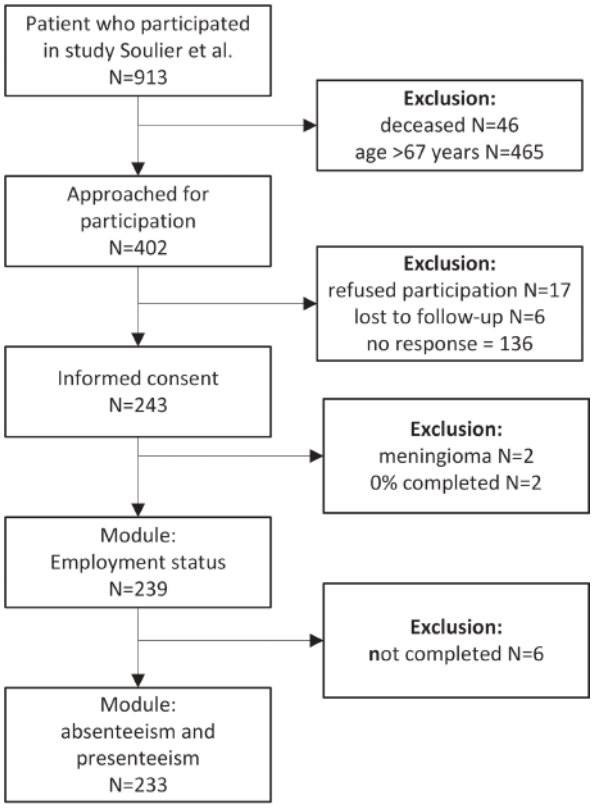


Figure 1. Flowchart
Patients who participated in a survey study in 2014 were reproached for participation in this study. Patients aged >67 years (retirement age in the Netherlands) were excluded. Two patients had a different diagnosis after pathology and were excluded. yrs.= years

As shown in Table 1, the study population consisted of 95 patients under active surveillance, 113 post-surgery patients, 24 post-radiotherapy and seven patients who underwent both surgery and radiotherapy over the years. Post-surgical patients were younger and had a larger tumor size at diagnosis compared to patients who underwent radiotherapy or surveillance, most likely a reflection of the indications for surgery in The Netherlands.

Table 1. Baseline characteristics.

	Total		Active surveillance		Surgery		Radiotherapy		Surgery and radiotherapy	
	N = 239		N=95		N=113		N=24		N=7	
	N	%	N	%	N	%	N	%	N	%
Age										
<45 yrs	14	6%	3	3%	9	8%	1	4%	1	14%
45-54 yrs	49	21%	18	19%	24	21%	4	17%	3	43%
55-64 yrs	135	57%	61	64%	58	51%	15	63%	1	14%
65-68 yrs	41	17%	13	14%	22	20%	4	17%	2	29%
Sex (male)	125	52%	55	58%	54	48%	12	50%	4	57%
Educational level										
low	60	25%	21	22%	30	27%	6	25%	3	43%
middle	82	34%	36	38%	36	32%	7	29%	3	43%
high	97	41%	38	40%	47	42%	11	46%	1	14%
Kanzaki at diagnosis										
intrameatal	73	31%	48	51%	17	15%	7	29%	1	14%
small (0-10mm)	55	23%	22	23%	26	23%	6	25%	1	14%
medium (11-20 mm)	59	25%	25	27%	25	22%	9	38%		
moderately large(21-30)	32	13%			28	25%	2	8%	2	29%
large (31-40 mm)	13	5%			11	10%			2	29%
giant (>40mm)	7	3%			6	5%			1	14%
Time in years (median)										
since treatment (IQR)	10	(8-12)	9	(8-11)	11	(9-14)	8	(6-9)	9	(7-13)

The baseline characteristics of all participants are shown. The right column shows the patient who underwent both surgery and radiotherapy since diagnosis. Kanzaki represents the classification of Kanzaki et al. of the tumor size at diagnosis. Yrs=years, IQR = interquartile range

Employment status

Figure 2 shows the employment status of the study population aged between 45-65 years (N= 196) and the Dutch population aged 45-65 years. Patients <45 years (N=14) and 66-67 (N=28) were excluded from this comparison. Unemployment and voluntary unemployment (e.g., housewife or husband) were comparable in both groups. VS patients were more often retired compared to the general Dutch population (χ2, p-value 0.006), although the size of the differences is rather small: retirement was found in 6% vs. 3%, respectively and disability pensions in 14% vs. 11%, respectively, resulting a slightly lower proportion of employment for VS patients (72% vs. 78%, respectively).

Unemployed VS patients were on average three years older, more often female (64% vs. 41%), and more likely to have a low level of education (37% vs. 19%) than employed patients. There were no differences between employed and unemployed patients for treatment modality or time since treatment.

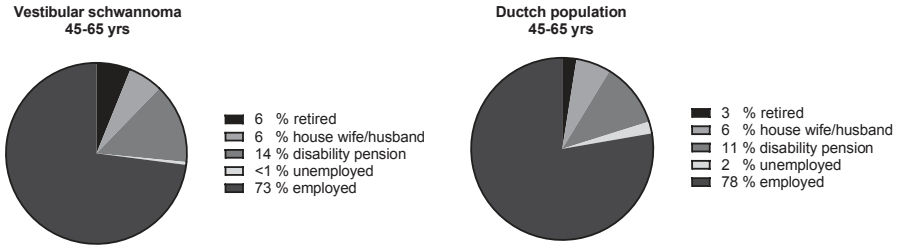


Figure 2. Employment status
The employment status of vestibular schwannoma patients aged 45-65 years (left) is compared to the reference population in the Netherlands (right). Patients who voluntarily do not have paid employment are labeled as 'house wife/husband'

The probability of being employed was assessed per treatment modality using logistic regression, as shown in Table 2. Model assumptions were met, although the explained variance of both models was relatively low. There were no differences between patients under active surveillance, after radiotherapy or after surgery. We found a tendency for a higher risk of unemployment in patients who underwent both surgery and radiotherapy. However, due to the small number of patients in this group, this was not statistically significant.

Table 2. Logistic regression assessing the effect of treatment modality on employment status (yes/no)

	Model 1[a]		Model 2[b]	
	OR	**CI95%**	**OR**	**CI95%**
Active surveillance (reference)	-		-	
Surgery	0.97	0.53;1.78	1.04	0.54;1.98
Radiotherapy	1.01	0.38;2.71	1.06	0.38;3.01
Surgery & radiotherapy	0.55	0.16;2.64	0.44	0.08;2.57
Sex (female)			**0.37**	**0.23;0.62**
Age			**0.83**	**0.81;0.86**
Educational level				
Low			0.66	0.36;1.21
Middle			1.02	0.58;1.79
High (reference)			-	
R^2	0.002		0.11	
χ^2	0.003		<0.001	

[a] treatment included. [b] treatment, sex, age and educational level included. OR= odds ratio, CI= confidence interval

Absenteeism and presenteeism

The number of working hours per week and the productivity losses due to absenteeism and presenteeism for employed VS patients (N=160) are shown in Table 3. Overall, the weekly working hours of VS patients did not differ significantly from the general Dutch employed population. In the active surveillance group, the working hours per week seemed slightly higher, but this difference disappeared after correction for confounding factors such as age, educational level and sex, as shown in the linear regression shown in Table 4.

Table 3. Absenteeism and presenteeism.

	Total	Active surveillance	Surgery	Radiotherapy	Dutch population[a]
	N=160	N=65	N=74	N=17	
Hours/week					
median (IQR)	36.0 (24-40)	36.0 (28-40)	32.0 (24-40)	36.0 (22-40)	
Mean (sd)	32.6 (12.1)	34.7 (10.4)	30.4 (10.5)	32.4 (14.9)	31
Difference pre diagnosis	-6.2%	-4.8%	-8.0%	-5.6%	
% presenteeism	1.8%	1.1%	2.6%	1.3%	
% absenteeism	4.2%	4.3%	4.3%	0%	4.4%

All patients we were employed are included in this table. Per treatment modality the working hours and productivity loss due to presenteeism and absenteeism are shown. There were four employed patients who underwent both surgery and radiotherapy, because of this small sample size they are not included as separate group in this table.[a] Dutch population statistics from 3th quarter of 2020 obtained from Statistics Netherlands.

Table 4. Linear regression assessing the effect of treatment modality on working hours per week

	Model 1[a]		Model 2[b]	
	Coefficient	95%CI	Coefficient	95%CI
Active surveillance (reference)	-		-	
Surgery	-4.0	-7.73;-0.35	-2.9	-6.17;0.29
Radiotherapy	-2.2	-8.13;3.37	-1.5	-6.68;3.70
Surgery & radiotherapy	4.1	-7.07;15.35	2.4	-7.64;12.47
Sex (female)			-11.33	-14.82;-7.85
Age			-0.21	-0.41;0.00
Educational level			0.22	-1.98;2.42
R^2	0.037		0.29	
Model ANOVA	0.12		<0.001	

[a] treatment included. [b]treatment, sex, age and educational level included. Educational level was categorized as 1,2,3, for low, middle and high level, respectively.

In addition to the working hours, the proportion of absence due to sick leave (absenteeism) was comparable to the sick leave rates in the general Dutch population, just over 4% of the total worked hours. Thirteen patients (8%) mentioned absenteeism in the last four weeks, causing an average of 17 hours of productivity loss in four weeks. Of these 13 patients, seven had been absent for the entire period of four weeks.

Furthermore, 23 patients (14%) mentioned that they were less productive while being at work. The extent of this presenteeism was low across all treatment modality groups, with a percentage of just under 2% of all worked hours. The productivity loss of these 23 patients were on average 3.2 hours per four weeks.

One out of six patients reported that their working hours have changed since the diagnosis. Of these 40 patients, 34 patients worked fewer hours and six worked more hours per week. Overall, the working hours decreased with 6%. Differences between treatment modality groups were statistically significant but small, with a decrease in working hours of 8%, 6% and 5% for surgery, radiotherapy and active surveillance, respectively ($\chi 2$, $p < 0.001$).

DISCUSSION

This study shows that employment status of patients with VS on average is quite comparable to a reference group of the general Dutch population. In addition, differences in sick leave rates (absenteeism) also were not statistically significant. Some patients however did report productivity losses while being at work (presenteeism). Employment rate, sick leave rate and productivity loss did not differ between the treatment modalities (surgery, radiotherapy and active surveillance), with the possible exception of patients receiving both radiotherapy and surgery.

Previous research on employment in VS patients described the differences pre- and post-treatment. Post-surgery, the percentage that could maintain their employment varied from 69-80%.[13,14,20] One of these studies reported that 79% of VS patients under active surveillance maintained their paid job.[20] Another small study on radiotherapy in young (<40yrs) VS patients reported that all patients maintained their employment.[12] In the current study, 15% of the patients reported a reduction in working hours after the diagnosis. The decrease in working hours (-8%) was the largest in the surgery group, followed by radiotherapy (-6%) and active surveillance (-5%). These differences are small and as the reduction was not necessarily due to illness or therapy related factors, causality is unclear. Even so, active treatment most likely will impact on working hours in

the recovery phase directly after treatment. For example, patients who undergo surgery are expected to have higher absenteeism rates in the direct postoperative phase. However, the results of this study indicate that this effect is temporary and that treatment modality does not differentially impact on long-term employment rates of VS patients, in contrast to age, sex and education level.

Employment status was compared with the Dutch population aged 45-65 years. This comparison showed that the proportion of disability pensions was almost similar. This finding contrasts with a Norwegian study, in which a threefold higher proportion of disability pensions (22%) was reported compared to the general population (6%).[11] This difference may be explained by the choice of reference population in the Norwegian study, since their reference group seemed to be the total Norwegian working-age population (i.e., all age groups). In our study, we opted for using the Dutch general population between 45-65 years as reference group because this matched the age distribution of the VS patients in this study most closely. As older people are more likely to have disability pensions, comparing with an age-matched population seems reasonable.

We found that the proportion of retired persons was mildly larger in the VS patient group than in the reference group. This difference is probably due to a slightly different age distribution in the reference and study populations, as patients aged 65 are relatively overrepresented in the latter. However, it is also conceivable that patients opt for early retirement due to sequelae of VS or its treatment.

Absenteeism (i.e., the number of working hours lost due to illness) in VS patients was similar to the entire Dutch working-age population in the third quarter of 2020, when the questionnaires were completed. Absenteeism was less prominent in the VS patients than in pituitary tumor or irritable bowel syndrome patients (8% vs. 40% vs. 34%). [21,22] For presenteeism, no reference from the general population was available. However, the presenteeism incidence was lower than in patients with pituitary tumors (14% vs. 39%) and irritable bowel syndrome (14% vs. 61%).

This study has some inherent limitations. The cross-sectional design of the study precludes causal inferences. In addition, radiotherapy was underrepresented in the study and below the required number needed for a power of 80%. This could lead to type II errors in which real differences in employment after radiotherapy were not identified. Furthermore, patients were asked about their employment situation before the diagnosis. As a consequence of the extended follow-up, the time of diagnosis exceeded the recommended recall period of 4 weeks ago. This prolonged period yields a risk of recall bias and the results of employment pre and post diagnosis should be interpreted with

care. Last, the study was performed in the Netherlands, a country with an extensive social security system. It might be possible that employment rates are lower than in countries with a more limited social security system. The setting of this study should be considered when translating the results to other countries.

Strengths of the study include the relatively large cohort of VS patients and their long-term follow-up (median 10 years). In addition, all three treatment modalities were included allowing analysis of their differential effect on both employment status and productivity. Furthermore, this is the first study that assessed absenteeism and presenteeism in VS patients.

CONCLUSION

This study suggests that long-term employment in VS patients on average is comparable to the employment in the general population, regardless of the treatment modality. There were no differences between sick leave and disability rates of VS patients and the age-matched general population. Although absenteeism is variable in VS patients and increased absenteeism may be expected shortly after active VS treatment, the results of this study indicate that the long-term prospects for the employment of VS patients in general are encouraging, irrespective of the treatment strategy. This information is valuable in counseling and medical decision making.

REFERENCES

1. Matthies C, Samii M (1997) Management of 1000 vestibular schwannomas (acoustic neuromas): clinical presentation. Neurosurgery 40:1-9; discussion 9-10
2. Vogel JJ, Godefroy WP, van der Mey AGL, le Cessie S, Kaptein AA (2008) Illness perceptions, coping, and quality of life in vestibular schwannoma patients at diagnosis. Otol Neurotol 29 (6):839-845. doi:DOI 10.1097/MAO.0b013e3181820246
3. Carlson ML, Link MJ, Wanna GB, Driscoll CLW (2015) Management of Sporadic Vestibular Schwannoma. Otolaryngologic Clinics of North America, vol 48. doi:10.1016/j.otc.2015.02.003
4. Bowling A (1995) What things are important in people's lives? A survey of the public's judgements to inform scales of health related quality of life. Soc Sci Med 41 (10):1447-1462. doi:10.1016/0277-9536(95)00113-l
5. Breivik CN, Varughese JK, Wentzel-Larsen T, Vassbotn F, Lund-Johansen M (2012) Conservative Management of Vestibular Schwannoma—A Prospective Cohort Study: Treatment, Symptoms, and Quality of Life. Neurosurgery 70 (5):1072-1080. doi:10.1227/neu.0b013e31823f5afa
6. van Leeuwen BM, Herruer JM, Putter H, Jansen JC, van der Mey AG, Kaptein AA (2013) Validating the Penn Acoustic Neuroma Quality Of Life Scale in a sample of Dutch patients recently diagnosed with vestibular schwannoma. Otol Neurotol 34 (5):952-957. doi:10.1097/MAO.0b013e31828bb2bb
7. Jahoda M (1982) Employment and unemployment: A social-psychological analysis, vol 1. Cambridge University Press,
8. Bartley M (2004) Employment status, employment conditions, and limiting illness: prospective evidence from the British household panel survey 1991-2001. J Epidemiol Community Health 58 (6):501-506. doi:10.1136/jech.2003.009878
9. Montgomery SM, Cook DG, Bartley MJ, Wadsworth ME (1999) Unemployment pre-dates symptoms of depression and anxiety resulting in medical consultation in young men. Int J Epidemiol 28 (1):95-100. doi:10.1093/ije/28.1.95 %J International Journal of Epidemiology
10. Stauder J (2019) Unemployment, unemployment duration, and health: selection or causation? The European Journal of Health Economics 20 (1):59-73. doi:10.1007/s10198-018-0982-2
11. Breivik CN, Nilsen RM, Myrseth E, Finnkirk MK, Lund-Johansen M (2013) Working disability in Norwegian patients with vestibular schwannoma: vertigo predicts future dependence. World Neurosurg 80 (6):e301-e305. doi:10.1016/j.wneu.2013.03.069
12. Lobato-Polo J, Kondziolka D, Zorro O, Kano H, Flickinger JC, Lunsford LD (2009) Gamma knife radiosurgery in younger patients with vestibular schwannoma. Neurosurgery 65 (2):294-301. doi:10.1227/01.neu.0000345944.14065.35
13. Nikolopoulos TP, Johnson I, O'Donoghue GM (1998) Quality of life after acoustic neuroma surgery. The Laryngoscope 108 (9):1382-1385. doi:10.1097/00005537-199809000-00024
14. Van Leeuwen JPPM, Meijer H, Braspenning JCC, Cremers WRJ (1996) Quality of Life after Acoustic Neuroma Surgery. Ann Otol Rhinol Laryngol 105 (6):423-430. doi:10.1177/000348949610500602
15. Neve OM, Soulier G, Hendriksma M, Van Der Mey AGL, Van Linge A, Van Benthem PPG, Hensen EF, Stiggelbout AM (2020) Patient-reported factors that influence the vestibular schwannoma treatment decision: a qualitative study. Eur Arch Otorhinolaryngol. doi:10.1007/s00405-020-06401-0
16. Bouwmans C, Krol M, Severens H, Koopmanschap M, Brouwer W, Roijen LH-V (2015) The iMTA Productivity Cost Questionnaire. Value Health 18 (6):753-758. doi:10.1016/j.jval.2015.05.009
17. Kanzaki J, Tos M, Sanna M, Moffat DA, Monsell EM, Berliner KI (2003) New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma. Otol Neurotol 24 (4):642-648; discussion 648-649. doi:10.1097/00129492-200307000-00019

18. Statistics Netherlands (2017) Standaard Onderwijsindeling 2016. Den Haag
19. Statistics Netherlands (2020) Labour participation. Statistics Netherlands,. https://opendata.cbs.nl/statline/#/CBS/en/dataset/82309ENG/table?ts=1614606737892. Accessed 10-02-2021 2021
20. Tos T, Caye-Thomasen P, Stangerup S-E, Tos M, Thomsen J (2003) Long-term socio-economic impact of vestibular schwannoma for patients under observation and after surgery. The Journal of Laryngology & Otology 117 (12):955-964. doi:10.1258/002221503322683830
21. Lobatto DJ, Steffens ANV, Zamanipoor Najafabadi AH, Andela CD, Pereira AM, Van Den Hout WB, Peul WC, Vliet Vlieland TPM, Biermasz NR, Van Furth WR (2018) Work disability and its determinants in patients with pituitary tumor-related disease. Pituitary 21 (6):593-604. doi:10.1007/s11102-018-0913-3
22. Weerts ZZRM, Vork L, Mujagic Z, Keszthelyi D, Hesselink MAM, Kruimel J, Leue C, Muris JWM, Jonkers DMAE, Masclee AAM (2019) Reduction in IBS symptom severity is not paralleled by improvement in quality of life in patients with irritable bowel syndrome. Neurogastroenterol Motil 31 (8). doi:10.1111/nmo.13629

Response rate of patient reported outcomes: the delivery method matters

Olaf Neve
Peter Paul van Benthem
Anne Stiggelbout
Erik Hensen

BMC Med Res Methodol. 2021 Oct 22;21(1):220
doi: 10.1186/s12874-021-01419-2

ABSTRACT

Background

Patient Reported Outcomes (PROs) are subjective outcomes of disease and/or treatment in clinical research. For effective evaluations of PROs, high response rates are crucial. This study assessed the impact of the delivery method on the patients' response rate.

Methods

A cohort of patients with a unilateral vestibular schwannoma (a condition with substantial impact on quality of life, requiring prolonged follow-up) was assigned to three delivery methods: email, regular mail, and hybrid. Patients were matched for age and time since the last visit to the outpatient clinic. The primary outcome was the response rate, determinants other than delivery mode were age, education and time since the last consultation. In addition, the effect of a second reminder by telephone was evaluated.

Results

In total 602 patients participated in this study. The response rates for delivery by email, hybrid, and mail were 45%, 58% and 60%, respectively. The response rates increased after a reminder by telephone to 62%, 67% and 64%, respectively. A lower response rate was associated with lower level of education and longer time interval since last outpatient clinic visit.

Conclusion

The response rate for PRO varies by delivery method. PRO surveys by regular mail yield the highest response rate, followed by hybrid and email delivery methods. Hybrid delivery combines good response rates with the ease of digitally returned questionnaires.

BACKGROUND

Patient Reported Outcomes (PROs) are increasingly used both for scientific purposes and in clinical practice. PROs measure the patients' perceived symptoms, functioning, and health-related quality of life. The use of PROs in research improves understanding the patient's perspective on the disease, the sequelae, and therapy.[1] In addition, using PROs in clinical practice may improve patient-clinician communication and enhance patient outcomes.[2, 3] However, the implementation of PROs in routine practice can be challenging due to technological and workflow barriers.[2]

One such barrier can be the response rate. A low response rate can lead to the introduction of selection bias and reduce the outcomes' external validity.[4] In general, response rates can be improved by several methods including monetary incentives, shorter questionnaires, reminders, personally addressed invitations and delivery method.[5-8] Delivery by email is increasingly used, with both distribution and digital data entry of the answers saving costs. However, delivery by regular mail has seemed to provide better response rates over the years.[8] Research performed in the medical context has shown that clinicians' response rates are similar or slightly in favor of mail delivery compared to email.[9, 10] A hybrid delivery method using both mail and email might be better than either email or mail alone.[11] Research on delivery method and patients' response rates is scarce and often performed in small sample sizes. These studies, published between 2014 and 2017, have shown that mail delivery results in higher response rates compared to email delivery.[12-14] However, digital literacy has rapidly increased in recent years. For example, in Europe 87% of the people aged 16-74 years had used internet in the last three months in 2019 compared to 75% in 2013, and 57% in 2007.[15] As a result, patients' response to email may have increased too. This study assessed three different delivery methods for PRO measures in a large cohort of patients with unilateral vestibular schwannoma.

METHODS

This study was part of a larger study on long-term outcomes of vestibular schwannoma management. Vestibular schwannoma is a benign, usually not life-threatening intracranial tumor, causing symptoms such as hearing loss, tinnitus, and balance problems due to pressure on adjacent structures, and as such may have considerable impact on quality of life. A small majority of these tumors is non-progressive and in these cases active surveillance during an extended follow-up period is usually the management option of choice. In progressive tumors, surgery or radiotherapy is performed to prevent

future complications such as brain stem compression or elevated intracranial pressure. After an active intervention, prolonged active surveillance ensues in these patients too, in order to identify possible recurrences.

Patients who participated in a survey study in 2014 were re-approached for participation in a survey between May and September 2020.[16] Both studies were performed at the Leiden University Medical Center, an expert referral center for vestibular schwannoma in the Netherlands. All patients were diagnosed with unilateral VS between 2003-2014. Patients with bilateral VS, other skull base pathologies or insufficient proficiency in the Dutch language to complete the questionnaires were excluded.

Several PRO measures that are used in the routine care for vestibular schwannoma care in our hospital were collected in this study. Patients received a general health-related quality of life (HRQL) questionnaire, the short form 36 (SF-36), and a disease-specific HRQL questionnaire, the Penn Acoustic Neuroma Quality-of-Life Scale (PANQOL).[17, 18] In addition, patients were asked to complete the dizziness handicap inventory (DHI), the medical outcome study cognitive functioning scale (MOS-CFS), the decision regret scale and the productivity costs questionnaires (iPCQ).[19-21] Combined, patients were asked to answer 117 questions.

Three different delivery methods were used: email, regular mail, and a hybrid of the two. These three methods were chosen because they represented the modern delivery method (email), the golden standard so far (mail) and an intermediate (hybrid) method that combines the conventional approach of mail with the advantage of digital data entry. Patients in the email group received an email invitation with a link to a digital informed consent form. After providing consent, patients were directed to digital questionnaires. Patients in the hybrid group were invited by regular mail with a letter including a unique code and a link to the digital informed consent form and the questionnaires. The regular mail group received an informed consent form, the printed questionnaires, and a prepaid return envelope. After two weeks, patients received a first personally addressed reminder by email (email group) or mail (hybrid and regular mail group). After another two weeks, all non-responders were called once by telephone for a second reminder. This telephone call was performed by a researcher, not their treating physician. In all groups, patients could request a different delivery method. Responders were defined as patients who completed the informed consent form and opened the questionnaire.

Before introducing electronic patient records in 2011, the patients' email address was not registered during the first visit to the hospital. Therefore, an email address was available for a minority of the patients, making randomization impossible. Patients for

whom the email address was registered were assigned to the email group. Patients from whom no email address was available were randomly assigned to either the regular mail or hybrid delivery groups. Two factors, age and time since the last visit, were expected to differ between groups with and without email, since most patients without email addresses were diagnosed before 2011. To avoid confounding of the effect of the delivery method on the response rate by two factors, we matched patients in all groups for age (<45yrs; 46-50yrs;…;81-85yrs;>85yrs) and time since the last visit (<5yrs;5-10yrs;>10yrs), as is shown in Figure 1.

The frequencies of categorical variables and means of numerical variables were calculated. Demographics of responders and non-responders were compared. Next, three analyses were performed because patients could switch delivery methods. First, a stringent analysis was performed in which switchers were considered as non-responders. Second, an intention to treat analysis was conducted in which patients were analyzed in their predefined delivery method. Third, an as treated analysis was performed in which patients who switched between delivery methods were analyzed in that category. The outcome was the response rate per group, which was analyzed using a chi-squared test. We also assessed the effect of the second telephone call reminder by a chi-squared test. In addition, the effect on the response rate of age, sex, education level, the time elapsed

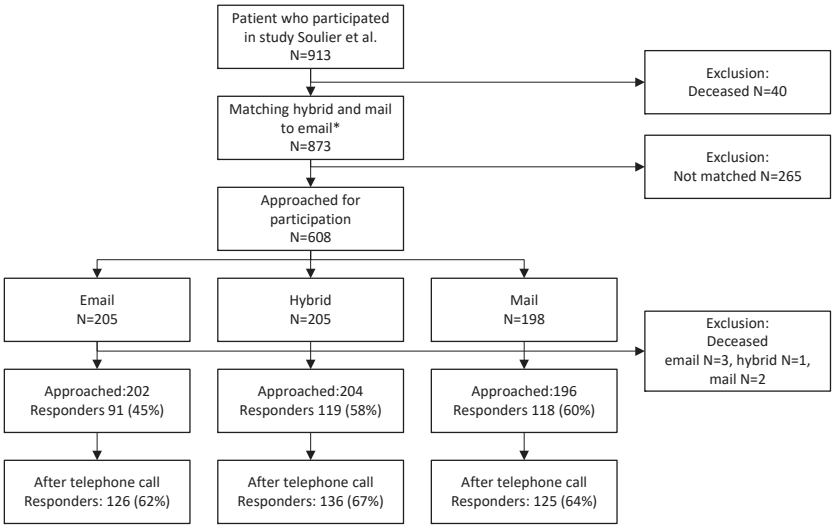


Figure 1. Flowchart of study participants.
Patients who participated in a previous study in 2014 were reapproached for participation. Before 2011 email was not registered at the first visit to the hospital. As a result, an email was available for a minority of the patients making randomization impossible. * All groups were matched for age and the time since the last visit.

since the last visit (in years), and the delivery method were analyzed using logistic regression with response rate as the dependent variable. The independent variables were selected based on their reported effect on response rates in previous literature.[8, 14, 22] Furthermore, interactions between independent variables were checked and, when relevant, included in the model. Model assumptions for multicollinearity were checked by calculating the variance inflation factor (VIF) and goodness of fit was verified with a Hosmer Lemeshow test and model chi-squared test. A minimum sample size of 387 was required based on a power calculation for the primary outcome, which used the difference in response rates in previous research (effect size w=0.2, α=0.05, 1-β=0.95).

All statistical analyses were performed in SPSS version 26 (Armonk, NY: IBM Corp). A p-value <0.05 was considered statistically significant. Demographic information was available from a previous 2014 study, so there were no missing data for any demographic variables.

RESULTS

In total, 602 patients were approached, of which 45 (7%) refused participation, 170 (28%) did not respond, and 387 (64%) responded, as is shown in Figure 1. Baseline characteristics of the patients in the three groups are shown in Table 1. As expected, the matching variables age and time elapsed since the last visit were equally distributed in all groups. The proportion of patients with a low level of education was larger in the mail group. Patients with a lower educational level, aged between 50-59 years or >80 years, or >5 years since the last visit were more often non-responders (Table 2).

Only 15 patients (2%) completed fewer than 80% of the total number of questions. Most incomplete responders in the email (N=5) and hybrid (N=6) groups seemed to have started the questionnaires and stopped at some point, without skipping items. In the mail group, incomplete responders (N=4) skipped some questions. Because of the low number of incomplete responders, statistical analysis of differences in item or PRO level response rates or differences per PRO questionnaire could not be reliably performed.
Furthermore, 98 (16%) patients used the possibility to request a different delivery method: 74 (76%) preferred to receive a questionnaire by regular mail and 24 (24%) preferred to complete the questionnaire electronically (by email).

Figure 2 shows the results of the three performed analyses. In the stringent analysis, mail delivery resulted in statistically significantly better response rates compared to email and hybrid 57% versus 37% and 38%, respectively (χ2, p<0.001). In the intention

Table 1. Baseline characteristics and response rates.

	Stringent/intention to treat			As treated		
	Email (N=202)	**Hybrid** (N=204)	**Mail** (N=196)	**Email** (N=201)	**Hybrid** (N=151)	**Mail** (N=250)
Sex (female)	98 (49%)	107 (53%)	91 (46%)	97 (48%)	76 (50%)	123 (49%)
Age						
<50 yrs	11 (5%)	11 (5%)	9 (5%)	10 (5%)	11 (7%)	10 (4%)
50-59 yrs	40 (20%)	40 (20%)	38 (20%)	39 (19%)	35 (23%)	44 (18%)
60-69 yrs	62 (31%)	63 (31%)	61 (31%)	67 (33%)	53 (35%)	66 (26%)
70-79 yrs	68 (34%)	70 (34%)	67 (34%)	60 (30%)	41 (27%)	104 (42%)
>79 yrs	21 (10%)	20 (10%)	21 (11%)	25 (12%)	11 (7%)	26 (10%)
Education level						
Low	55 (27%)	78 (38%)	84 (43%)	64 (32%)	48 (32%)	105 (42%)
Middle	59 (29%)	52 (26%)	51 (26%)	57 (28%)	42 (28%)	63 (25%)
High	88 (44%)	72 (35%)	60 (31%)	80 (40%)	60 (40%)	80 (32%)
Time since last visit						
<5 yrs	126 (62%)	125 (61%)	120 (61%)	122 (60%)	91 (60%)	160 (64%)
≥5 yrs	76 (38%)	79 (39%)	76 (39%)	81 (40%)	60 (40%)	90 (36%)
Response rate						
Stringent	75 (37%)	77 (38%)	112 (57%)			
After 1st reminder	91 (45%)	119 (58%)	118 (60%)	85 (42%)	77 (51%)	166 (66%)
After telephone call	126 (62%)	136 (67%)	125 (64%)	120 (60%)	94 (62%)	173 (69%)
Different delivery method						
Email	-	4	20			
Mail	25	49	-			

Baseline characteristics of the three delivery methods for the stringent, intention to treat and as treated analysis are shown. The stringent response rate considered patients who switched delivery method as non-responders. In the intention to treat analysis patients are grouped in the delivery method category they were assigned to. In the as treated patients were grouped their actual delivery method category, since some patients had requested a different delivery method. Time since the last visit shows the years since the last consultation in the hospital. yrs=years

to treat analysis, when patients who switched delivery method were included, the response rates for patients allocated to delivery by email, hybrid, and regular mail were 45%, 58% and 60%, respectively ($\chi 2$ $p<0.001$).

The requests for a different delivery method resulted in a decrease in email (-0.5%; N=-1) and hybrid delivery (-26%; N=-53), and an increase in mail delivery (+28%; N=+54), as is shown in Table 1. The response rate for the actual delivery method, shown in the as treated analysis, was 42% by email, 51% by hybrid, and 66% by mail ($\chi 2$, $p<0.001$).

Table 2. Non-responder analysis.

	Responders	Non-responder	% responder
N	387	215	64%
Sex			
Female	187 (48%)	109 (51%)	63%
Male	200 (52%)	106 (49%)	65%
Age			
<50 yrs	20 (5%)	11 (5%)	64%
50-59 yrs	66 (17%)	52 (24%)	56%
60-69 yrs	125 (32%)	61 (28%)	67%
70-79 yrs	143 (37%)	62 (29%)	70%
>79 yrs	33 (9%)	29 (14%)	53%
Education level			
Low	128 (33%)	89 (41%)	59%
Middle	109 (28%)	53 (25%)	67%
High	148 (38%)	72 (34%)	67%
Time since the last visit			
<5 yrs	254 (66%)	117 (54%)	69%
≥5 yrs	133 (34%)	98 (46%)	58%

The demographics of overall responders (after first and second reminder) compared to non-responders. The percentages in the second and third columns reflect the percentage within the demographic group. The last column, % responder, reflects the percentage responders of each variable.

Table 3. Effect of telephone call reminder.

	Responders	Non-responders		
		Not answered	Non-responder despite promise	Refused to participate
N (% of total)	59 (24%)	125 (50%)	45 (18%)	19 (8%)
Sex (female)	32 (54%)	66 (53%)	23(50%)	11 (58%)
Mean age (sd)	65.9 (11.9)	65.9 (11.1)	63.8 (11.4)	71 (11.3)
Education level				
Low	24 (41%)	55 (44%)	13 (29%)	11 (58%)
Middle	20 (34%)	29 (23%)	16 (36%)	3 (16%)
High	15 (41%)	41 (33%)	16 (36%)	5 (26%)
Time since the last visit				
<5 yrs	40 (68%)	66 (53%)	23 (51%)	13 (68%)
≥5 yrs	19 (32%)	59 (47%)	22 (49%)	6 (32%)

All non-responders (N=248) were called two weeks after the first reminder. This table shows the demographics of this group. Half of the patients did not answer the telephone call. When patients did answer the telephone 59 out of 123 did participate, while 19 refused to participate. Another 45 patients promised to participate on the telephone but did not participate eventually.

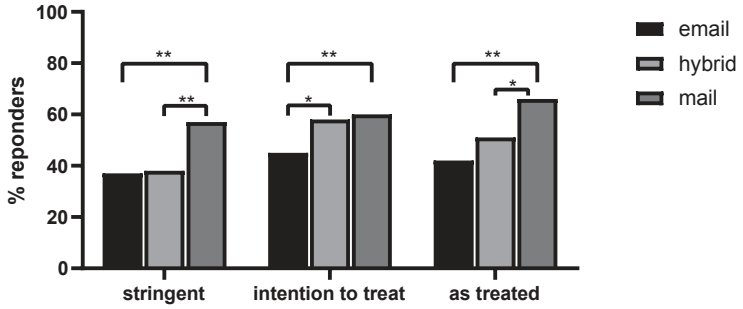


Figure 2. Response rates.
The response rates of the different delivery methods are shown per analysis. In the stringent analysis, patients who requested a different delivery method are considered non-responders. In the intention to treat analysis all patients are analysed in their predefined group and in the as treated in their actual delivery method. * = $\chi 2$ p-value <0.01. **= $\chi 2$ p-value <0.001

Reminder by telephone

After the first reminder by either email or regular mail, 248 patients (41%) still did not respond and received a reminder by telephone call. Nearly half of these (N=123) initial non-responders answered the telephone, of whom 48% (N=59) did participate after this telephone call, 36% (N=45) did not respond while they said to do so in the telephone call, and 15% (N=19) declined participation. The demographics of these groups are shown in Table 3. The response rates in the intention to treat analysis raised to 62%, 67%, and 64% for email, hybrid and mail, respectively ($\chi 2$ p=0.65). In the as treated analysis the final response rates were 60%, 62% and 69%, respectively ($\chi 2$, p=0.09).

Logistic regression

The results of the logistic regression are shown in Table 4. The stringent, intention to treat and as treated models met the model assumptions and goodness of fit tests. All models showed that the probability of responding was lower in the email delivery group. The hybrid delivery was also associated with a lower response rate in the stringent and the as treated models.

A low education level was a confounding factor in all models. Age and sex did not contribute to a lower or higher response rate, except for patients aged 60-69 years in the stringent model, who were more likely to respond.

The interaction between the time since the last visit and delivery method was close to statistical significance in the intention to treat and as treated analyses. In the stringent analysis, this interaction was statistically significant, meaning that patients whose last visit to the hospital was less than five years ago tended to have different response rates per delivery method than patients whose last visit was longer ago. In the mail delivery group, the response rate decreased with increasing time since last visit. In the other

Table 4. Results of logistic regression.

		Stringent		Intention to treat		As treated		
		OR	**95%CI**	**OR**	**95% CI**		**OR**	**95% CI**
Delivery method								
Email	N=202	*0.24*	*(0.14-0.41)*	*0.32*	*(0.21-0.60)*	*N=201*	*0.27*	*(0.16-0.45)*
Hybrid	N=204	*0.34*	*(0.22-0.54)*	0.77	(0.49-1.21)	N=151	*0.45*	*(0.28-0.71)*
Mail (reference)	N=196	-		-		N=250	-	
Sex (female reference)		1.16	(0.82-1.64)	1.08	(0.77-1.52)		1.11	(0.79-1.56)
Age								
<50 yrs		1.20	(0.47-3.08)	0.79	(0.32-1.96)		0.81	(0.32-2.03)
50-59 yrs		1.54	(0.77-3.07)	1.16	(0.60-2.23)		1.13	(0.58-2.18)
60-69 yrs		2.16	(1.13-4.01)	1.49	(0.81-2.72)		1.48	(0.80-2.74)
70-79 yrs		1.43	(0.76-2.69)	1.69	(0.94-3.06)		1.49	(0.81-2.71)
>79 yrs (reference)		-		-			-	
Education level								
Low		*0.45*	*(0.29-0.69)*	*0.47*	*(0.31-0.72)*		*0.48*	*(0.31-0.73)*
Middle		0.87	(0.57-1.33)	0.75	(0.48-1.14)		0.76	(0.50-1.17)
High (reference)		-		-			-	
Time since the last visit								
<5 yrs		0.45	(0.18-1.12)	0.68	(0.28-1.72)		0.77	(0.31-1.87)
≥5 yrs (reference)		-		-			-	
Interaction term								
time last visit *delivery method		*0.55*	*(0.36-0.84)*	0.67	(0.43-1.03)		0.72	(0.48-1.08)
Model χ^2		<0.001		<0.001			<0.001	
Hosmer and Lemeshow		0.43		0.51			0.18	

In all regression models, response rate was the dependent variable. The independent variables were delivery method, age, sex, education level, and time since the last visit. Also, the interaction term time since the last visit * delivery method was included; other interaction terms were not significant and therefore not included in the regression models. The χ^2 and the Hosmer and Lemeshow show the p-values of the goodness of fit tests. The time since the last visit is the years since the last consultation in the hospital. Statistically significant effects are shown in italics. OR=Odds ratio, CI= confidence interval, yrs=years

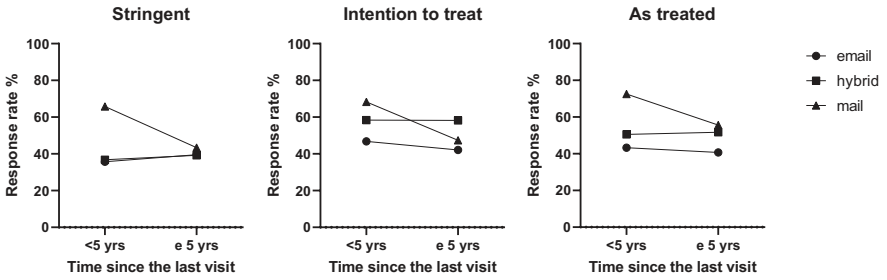


Figure 3. Interaction delivery method and time since the last visit.
The time since the last visit affected the relation between delivery method and the response rate. In all analyses (stringent, intention to treat, as treated), response rate decreased with increasing time since last visit. This effect was not observed in the email and hybrid delivery groups

groups, this effect was not observed, as is shown in figure 3. Other interactions (i.e. between age, sex, education level, and delivery method) were not statistically significant (lenient p-values of more than 0.2) and were not included in the models.

DISCUSSION

This study suggests that email delivery might result in a lower response rate compared to delivery by regular mail or hybrid delivery. Even when patients could choose their preferred delivery method, the response rate per email remained lower than mail or hybrid delivery.

The low response rate of email delivery is consistent with prior studies on patient response.[12, 14] This is somewhat surprising as one might expect increasing digital literacy in patients with the growing digitalization of the patient journey in hospitals today. Compared to other studies, we found smaller differences between the delivery methods, despite patients' older average age in this study. An older population might be less familiar with the internet or email, but in The Netherlands, 87% of the elderly (>65 years) have internet access, and 72% used email in 2019. In the subgroup of 65-75 years (which comprises approximately half our study population), these percentages are even higher: 95% internet access and 83% use of email.[23]

Sex and education level could also act as confounding factors factors on response rate or interact with delivery method. For example, in healthcare-related research amongst patients, an effect of sex is not consistently observed.[24, 25] In this study too, sex did not seem to affect the response rate or vary the response rate by delivery method (i.e., no significant interaction with delivery method). The level of education did have a signifi-

cant impact on response rates, as patients with a low level of education were less likely to be responders in this study (table 4), which is consistent with a previous report.[8] However, the effect of the delivery method on response rate did not vary by education level. Finally, the time since last clinic visit appeared to affect the association between delivery method and response rate, as we observed a decreasing response rate with increasing time since last visit, but only in the mail delivery group (figure 3). This effect might be comparable to the effect of decreasing response rates with increasing follow-up periods, as reported in long-term follow-up studies, however it is unclear why this effect is only seen after mail delivery.[26]

Although regular mail delivery had the highest response rate, there are some logistic disadvantages. To use the PROs, surveys on paper need to be digitized, which is time-consuming and error-prone. This is especially cumbersome when PROs are used in a clinical context, and feedback is expected during clinical consultation. In this light, the results of hybrid delivery are noteworthy since the response rate is close to regular mail delivery, but the PROs are completed and returned electronically. In practice, using a hybrid system could reduce the workload of digitizing PRO outcomes, with comparable response rates to surveys by mail.

In addition, a telephone call reminder can further increase response rates. In the current study, 48% of initial non-responders did respond after being reminded by a telephone call. However, the advantage of this higher response rate should be weighed against the time investment needed.

There are some inherent limitations to this study. First, it was impossible to perform a randomized trial because an email address was not available for all patients eligible for inclusion. Although the missing email addresses were caused by a different registration system in the hospital, we cannot be entirely sure that the differences between the groups are purely random. Second, the study participants were probably prone to participate in a research survey because they had already participated in a previous study in 2014. This committed population may therefore have increased response rates. Conversely, a decreased response rate may have been caused by a prolonged time interval between the survey and the last consultation, as was observed in a number of participants and was associated with a lower probability of responding in this study. Last, the PRO measures response rates found in this cross-sectional research setting may not be representative of PRO measures response rates in a clinical setting, in which PRO measures are typically collected close before or after a clinical consultation and serve a more direct clinical purpose. However, patient preferences with regard to the survey delivery method are probably equally applicable to both settings.

When using PRO measures, the response rate is an essential factor to consider. Various factors have been identified that influence the response rate, such as personally addressed invitations, shorter questionnaires, and financial incentives.[7, 27, 28] In the current study, all invitations were personally addressed, but no financial incentives or differences in questionnaire lengths were applied. In addition, we found that a reminder by letter and/or telephone call may be a particularly important factor in increasing the response rate of patients, which is in agreement with previous report on health studies.[7] In addition, this study suggests that two other factors are of importance in patients' response rates: the initial delivery method and the ability to choose the desired delivery method.

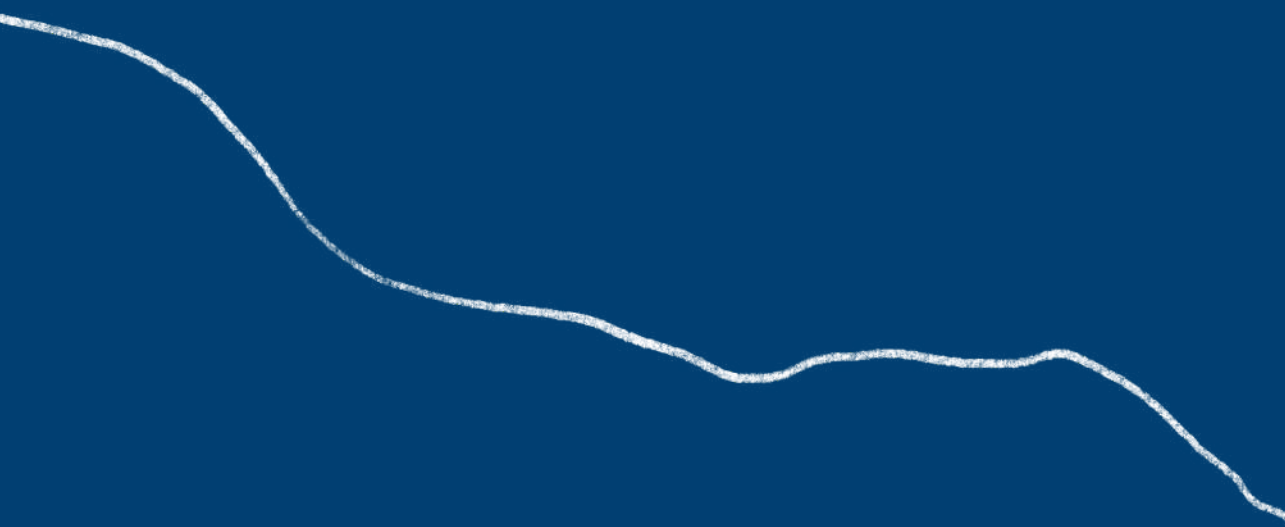
CONCLUSION

The effectiveness of the increasing use of PROs in healthcare stands or falls by patients completing and returning the questionnaires. This response rate can be influenced by several aspects, and the current study suggests that the route of survey delivery is an important factor. Regular mail delivery seems to perform better than email delivery in our study population but is more time-consuming, both in distribution, and in digitalization afterwards. Therefore, a hybrid delivery method in which patients receive a letter by regular mail with a code to access the survey digitally might be the best of both worlds.

REFERENCES

1. Basch E, Abernethy AP, Mullins CD, Reeve BB, Smith ML, Coons SJ, Sloan J, Wenzel K, Chauhan C, Eppard W et al: Recommendations for Incorporating Patient-Reported Outcomes Into Clinical Comparative Effectiveness Research in Adult Oncology. J Clin Oncol 2012, 30(34):4249-4255.
2. Basch E: Patient-Reported Outcomes — Harnessing Patients' Voices to Improve Clinical Care. N Engl J Med 2017, 376(2):105-108.
3. Kotronoulas G, Kearney N, Maguire R, Harrow A, Di Domenico D, Croy S, Macgillivray S: What Is the Value of the Routine Use of Patient-Reported Outcome Measures Toward Improvement of Patient Outcomes, Processes of Care, and Health Service Outcomes in Cancer Care? A Systematic Review of Controlled Trials. J Clin Oncol 2014, 32(14):1480-1501.
4. Johnson TP: Response Rates and Nonresponse Errors in Surveys. JAMA 2012, 307(17):1805.
5. Edwards PJ, Roberts I, Clarke MJ, Diguiseppi C, Wentz R, Kwan I, Cooper R, Felix LM, Pratap S: Methods to increase response to postal and electronic questionnaires. Cochrane Database Syst Rev 2009(3):Mr000008.
6. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R: Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017, 17.
7. Nakash RA, Hutton JL, Jørstad-Stein EC, Gates S, Lamb SE: Maximising response to postal questionnaires – A systematic review of randomised trials in health research. BMC Med Res Methodol 2006, 6(1):5.
8. Shih T-H, Fan X: Comparing response rates in e-mail and paper surveys: A meta-analysis. Educational Research Review 2009, 4(1):26-40.
9. Weaver L, Beebe TJ, Rockwood T: The impact of survey mode on the response rate in a survey of the factors that influence Minnesota physicians' disclosure practices. BMC Med Res Methodol 2019, 19(1):73.
10. Hardigan PC, Popovici I, Carvajal MJ: Response rate, response time, and economic costs of survey research: A randomized trial of practicing pharmacists. Research in Social and Administrative Pharmacy 2016, 12(1):141-148.
11. Beebe TJ, Jacobson RM, Jenkins SM, Lackore KA, Rutten LJF: Testing the Impact of Mixed-Mode Designs (Mail and Web) and Multiple Contact Attempts within Mode (Mail or Web) on Clinician Survey Response. Health Serv Res 2018, 53:3070-3083.
12. Palmen LN, Schrier JCM, Scholten R, Jansen JHW, Koëter S: Is it too early to move to full electronic PROM data collection? Foot Ankle Surg 2016, 22(1):46-49.
13. Feigelson HS, McMullen CK, Madrid S, Sterrett AT, Powers JD, Blum-Barnett E, Pawloski PA, Ziegenfuss JY, Quinn VP, Arterburn DE et al: Optimizing patient-reported outcome and risk factor reporting from cancer survivors: a randomized trial of four different survey methods among colorectal cancer survivors. J Cancer Surviv 2017, 11(3):393-400.
14. Nota SPFT, Strooker JA, Ring D: Differences in Response Rates between Mail, E-mail, and Telephone Follow-Up in Hand Surgery Research. Hand 2014, 9(4):504-510.
15. Eurostat: Individual Internet use. In. https://ec.europa.eu/eurostat/data/database: Eurostat; 2019.
16. Soulier G, Van Leeuwen BM, Putter H, Jansen JC, Malessy MJA, Van Benthem PPG, Van Der Mey AGL, Stiggelbout AM: Quality of Life in 807 Patients with Vestibular Schwannoma: Comparing Treatment Modalities. Otolaryngology–Head and Neck Surgery 2017, 157(1):92-98.

17. Ware JE, Jr., Sherbourne CD: The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992, 30(6):473-483.
18. Shaffer BT, Cohen MS, Bigelow DC, Ruckenstein MJ: Validation of a disease-specific quality-of-life instrument for acoustic neuroma. The Laryngoscope 2010, 120(8):1646-1654.
19. Jacobson GP, Newman CW: The Development of the Dizziness Handicap Inventory. Arch Otolaryngol Head Neck Surg 1990, 116(4):424-427.
20. Bouwmans C, Krol M, Severens H, Koopmanschap M, Brouwer W, Roijen LH-V: The iMTA Productivity Cost Questionnaire. Value Health 2015, 18(6):753-758.
21. Stewart A, Ware J, Sherbourne C, Wells K: Psychological distress/well-being and cognitive functioning measures In: Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. edn. Edited by Stewart A, Ware J. Durham, NC: Duke University Press; 1992: 102-142.
22. Matthews FE, Chatfield M, Freeman C, McCracken C, Brayne C, Cfas M: Attrition and bias in the MRC cognitive function and ageing study: an epidemiological investigation. BMC Public Health 2004, 4(1):12.
23. Statistics Netherlands: Internet; access and use. In. https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83429NED/table?fromstatweb; 2019.
24. Van Loon AJM, Tijhuis M, Picavet HSJ, Surtees PG, Ormel J: Survey non-response in the Netherlands: effects on prevalence estimates and associations. Ann Epidemiol 2003, 13(2):105-110.
25. Polk A, Rasmussen JV, Brorson S, Olsen BS: Reliability of patient-reported functional outcome in a joint replacement registry. Acta Orthop 2013, 84(1):12-17.
26. Westenberg RF, Nierich J, Lans J, Garg R, Eberlin KR, Chen NC: What Factors Are Associated With Response Rates for Long-term Follow-up Questionnaire Studies in Hand Surgery? Clinical Orthopaedics and Related Research® 2020, 478(12):2889-2898.
27. Edwards P: Increasing response rates to postal questionnaires: systematic review. BMJ 2002, 324(7347):1183-1183.
28. Vangeest JB, Johnson TP, Welch VL: Methodologies for Improving Response Rates in Surveys of Physicians. Eval Health Prof 2007, 30(4):303-321.



Patient reported factors that influence the vestibular schwannoma treatment decision: a qualitative study

Olaf Neve
Géke Soulier
Martine Hendriksma
Andel van der Mey
Anne van Linge
Peter Paul van Benthem
Erik Hensen
Anne Stiggelbout

Eur Arch Otorhinolaryngol. 2021 Sep;278(9):3237-3244.
doi: 10.1007/s00405-020-06401-0

ABSTRACT

Purpose

In cases of small- to medium-sized vestibular schwannomas, three management strategies can be opted for: active surveillance, surgery or radiotherapy. In these cases, the patient's preference is pivotal in decision making. The aim of this study was to identify factors that influence a patient's decision for a particular management strategy.

Methods

A qualitative inductive thematic analysis was performed based on semi-structured interviews. Eighteen patients with small- to medium-sized vestibular schwannomas were interviewed. All patients were diagnosed or treated at one of the two participating university medical centers in the Netherlands

Results

Ten themes were identified that influenced the decision, classified as either medical or patient-related. The medical themes that emerged were: tumor characteristics, the physician's recommendation, treatment outcomes and the perceived center's experience. The patient-related themes were: personal characteristics, anxiety, experiences, cognitions, logistics and trust in the physician.

Conclusion

Knowledge of the factors that influence decision making helps physicians to tailor their consultations in order to arrive at a true shared decision on vestibular schwannoma management.

INTRODUCTION

Vestibular schwannoma (VS) is a benign intracranial tumor, arising from schwann cells of the vestibular branch of the vestibulocochlear nerve. Current management options for VS generally consist of one of three modalities: microsurgery, radiotherapy or active surveillance (also known as wait-and-scan policy).[1] Based on the available evidence, a clinical equipoise exists in the management of small- to medium-sized VS (up to 25mm in extrameatal diameter). Active surveillance is a valid management option, especially in non-progressing tumors. However, hearing loss and vestibular problems may increase, even in otherwise stable tumors. In patients with (progressing) medium-sized tumors, both surgery and radiotherapy are viable options with equally high tumor control rates and generally good facial nerve outcome. Long-term hearing results are universally poor and vestibular function may be impacted, both after surgery and radiotherapy. However, both modalities differ considerably in their mode of action, administration and the way eventual sequelae and side effects become apparent, either suddenly (i.e., after surgery) or after a time interval (i.e., radiotherapy). The advantages and disadvantages of both modalities need to be weighed and patient and physician can explore treatment options together, a process that is also known as shared decision making (SDM). In case there is no clear superiority of one modality over the other from a medical point of view, patient preferences are important. SDM has been argued to be the preferred model in preference sensitive decisions, as in small- to medium-sized VS management.[2-4] A four-step model is often used to apply SDM in clinical practice.[4] Firstly, the physician informs the patient that a decision is to be made and that the patient's opinion is important. Secondly, the pros and cons of each relevant option are explained. Next, the physician and patient discuss the patient's preferences and the physician supports the patient in deliberation. Finally, the patient's decisional role and preference are discussed and the decision is made or deferred. To make this process work satisfactorily, an understanding of the factors that influence patients' decision are essential.

Known medical factors influencing decision making in VS are tumor size, tumor progression, symptoms, risk of complications and the physician's recommendation.[5] Practice variation may arise if patients do not receive unbiased information about all possible treatment options.[6,5] Patient factors that influence the VS treatment decision are less well known. As of yet, the only identified factors that influence decision making are anxiety and logistics.[7,8] For various other diseases, it has been reported that patient's coping and decision making style are also influential in the clinical decision. However, these factors have not been reported in VS patients.[9,10] This qualitative study aims to identify factors that influence a patient's decision.

MATERIALS AND METHODS

Design

A qualitative interview study was performed, followed by an inductive thematic analysis. Qualitative research allows exploring the notions of the respondents, without directing the answers by predefined questions or answering categories, as is the case in quantitative research. It can provide rich data and new insights about patient-reported factors of importance for treatment decisions.[11] Using this thematic qualitative analysis, patterns within the data were identified, analyzed and reported. Methods and results are reported in accordance with Standards of Reporting Qualitative Research.[12]

Ethics

The Medical Ethics Committee of the Leiden University Medical Center reviewed the protocol (P16.064) and concluded that their approval was not required under Dutch law.

Recruitment

Purposive sampling was used to enroll patients with medium-sized VS (i.e., 10 to 25 mm extrameatal diameter) from two tertiary care centers in the Netherlands, Erasmus Medical Center (EMC) and Leiden University Medical Center (LUMC). Both centers are experienced in the treatment of VS and offer surgery (mostly the translabyrinthine or retro sigmoid approach) and stereotactic fractionated (LUMC) or single dose (EMC) radiotherapy. Ambulatory procedures are similar, consisting of an initial consultation with an otorhinolaryngologist, followed by a multidisciplinary team (MDT) meeting attended by an otorhinolaryngologist, a neurosurgeon and a radiation oncologist and a subsequent consultation to discuss treatment options. Potential participants were identified from the records of the weekly MDT meeting in which all new patients with a VS are discussed. Patients were provided with information about the study by their treating physician during the subsequent consultation. One of the researchers (GS) followed up with a phone call after one week to check the willingness to participate.

Interviews

Semi-structured, face-to-face interviews were conducted to gather nuanced and context-dependent data.[13] The interviews were carried out using a topic guide, an outline of key issues and areas to explore during the interview (Table 1).This form of interviewing allows for new ideas to be brought up during the interview and to be incorporated in subsequent interviews. Participants were interviewed at a location of their choice (generally their own home). Conversations typically lasted between 30 and 90 minutes and were audio recorded. All interviews were conducted in Dutch and by the same interviewer (GS), who was not involved in patient care during this period.

Table 1.

Topic guide interviews
- Contextual information
- Information provision
- People of influence
- Aim of treatment
- Decision making process
- Priorities in decision making
- Barriers in decision making

Topic guide used for semi structured interviews

Interviewing was carried out until data saturation occurred, which was defined as the point when no new ideas emerged from the interviews. To this aim, data analysis was carried out concurrently. The stage at which data saturation occurred was determined by consensus within the research team (GS, MH and ON).

Analysis

All interviews were transcribed verbatim and imported in the qualitative analysis software ATLAS.ti (ATLAS.ti, version 8.4.18, GmbH; Berlin, Germany). Data were analyzed using the framework method. This method uses a framework matrix for data interpretation by charting in rows (patients) and columns (codes). This provides a structure into which data can be systematically reduced, facilitating analysis.[14,15] Data was coded with open codes that emerged from the text. Codes were initially assigned by one researcher

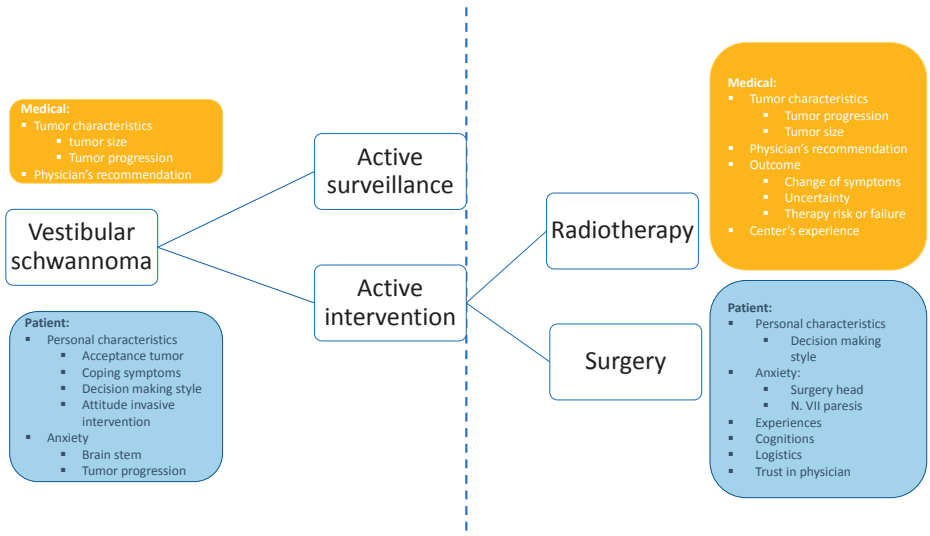


Figure 1. Two-step decision model based on the patient's experiences with factors that influence the decision making

(for the first eight interviews by GS and the other ten by ON). During the coding, both researchers met regularly with another member of the research team (MH) to review the codebook and discuss the interpretation of data. Coding was reviewed in ten interviews by a third researcher (MH). All research team members are medical doctors (MD). ON is a medical doctor and researcher trained in coding and analyzing patient interviews. At the time of the study, MH and GS were involved in patient care as trainee specialists. However, none of the researchers that conducted or analyzed the interviews (MH, GS or ON) were directly involved in the care of the study patients.

RESULTS

Data saturation was reached after 18 interviews. Nine patients from each center were included, one patient visited both centers. In eight interviews a spouse or a relative was present.

Patients experienced the decision making as challenging because it was hard to weigh the advantages and disadvantages of the treatment modalities.

> *"It is a choice between… Actually, between three evils. There is nothing good. None of the three is good, because all have their consequences" patient 2*

Decision making styles varied among the patients. Some patients indicated to defer decisions generally, this strategy usually resulted in a preference for active surveillance. Other patients were more decisive and had even decided before the first consultation.

> *"I can still postpone [the decision] a little, so I don't let it get to me any more than I need to at the moment" patient 9*

This difference in patients' decision making style was also reflected in the search for information on the disease and the treatment modalities. Some patients refused to search for information on the internet because of the fear of finding the worst case stories. A majority of the patients looked for information on websites of hospitals and patient associations. Others also looked for patient's experiences on social media. A minority searched in medical literature.

Although from a medical point of view, multiple management options and timing of possible interventions make the clinical decision complex, patients expressed that they experienced a rather straightforward two-step decisional process, as shown in Figure

1. Firstly, the decision between active surveillance and active treatment had to be made. Secondly, when active treatment was chosen, the patient had to decide between surgery and radiotherapy. Each decision was influenced by different factors, although a few factors influenced both decisions. We classified all factors as either medical or patient-related. Medical factors were defined as factors related to the tumor, physician or the treatment modality and patient factors were related to personal characteristics, experiences or cognitions.

Active surveillance vs. active treatment

The medical factors that patients perceived to have influenced the first decisions were tumor characteristics and physician's recommendation. Tumor characteristics, such as tumor size and progression, were important to determine whether active surveillance was still considered an option. Conversely, in the absence of tumor progression, active surveillance was generally deemed preferable.

> *"When the tumor is growing, surgery should be performed" patient 11*

In addition, the physician's recommendation affected decision making. Most patients acknowledged the authority of physicians because of their knowledge and experience.

> *"Physicians have influence on the decision. Despite all information on the internet nowadays, I rely on their know-how, they know what they are talking about, that is really important" patient 5*

Several patients were content when physicians offered all treatment options without a recommendation, however, some wanted more guidance from the physician. When no recommendation was given, these patients looked for clues as to the physician's own treatment preference. Some of them were surprised and sometimes disappointed when physicians did not provide any treatment advice. The majority of the patients, however, stated that in the end they felt that the decision was theirs to make, regardless of their need for guidance by their physician.

The decision between active surveillance or active treatment was influenced by two patient factors: anxiety and personal characteristics. Anxiety, specifically about tumor progression and brainstem compression, prompted patients to choose active treatment, all the more so if patients were surprised by the close anatomical relation between the tumor and the brainstem.

> *"It came as a shock … So, it [the tumor] is pushing against the brainstem." patient 7*

Personal characteristics included coping with symptoms, tumor acceptance, attitudes toward invasive treatment and decision making style. Patients that were less troubled by their symptoms tended to prefer active surveillance. Some patients preferred avoiding medical interventions in general and thus favored active surveillance. In contrast, some patients experienced difficulty with the concept of a persisting tumor inside their head. This lack of tumor acceptance made patients choose an active treatment.

> *"I want it [the tumor] to be removed! Surgery. Get rid of it. " patient 3*

Surgery vs. radiotherapy

Medical factors influencing the decision between surgery and radiotherapy were treatment outcome and a perceived center's experience with the treatment. In addition, this decision was also influenced by tumor characteristics and physician's recommendation. Treatment outcome contained three components: change of symptoms, uncertainty about tumor control and therapy risk or failure. Uncertainty about the occurrence of complications made the decision complex because patients found it hard to translate the group-based probabilities to their individual situation.

The perceived possibility of a change in symptoms or symptom severity as a result of therapy, such as the possibility of improvement of vestibular or trigeminal symptoms after surgery also affected patients' choices. Patients were generally aware that neither treatment modality improves hearing or facial paresis. Uncertainty about tumor control and therapy-associated complications influenced the patients' decision.

> *"At first, I looked whether there was a treatment that could improve my symptoms, there isn't. Then, I looked at the least risky treatment." patient 3*

Uncertainty about results after radiotherapy was caused by the lengthy time interval between the treatment and its effects or its complications. Because of this uncertainty, some patients preferred surgery, after which the effects and complications are immediately apparent.

> *"What also was of influence, is that we understood that radiotherapy… You never know immediately whether it was effective, it can take a year, in the worst case, to notice the effect or to notice that it has done nothing." Patient 9*

A proportion of the patients mentioned that the possible consequences of treatment failure influenced their decision making. Treatment failure was defined as tumor progression after initial treatment that necessitated additional active treatment. When

surgery had been performed as initial therapy, radiotherapy is the additional treatment of choice, if necessary and vice versa. Patients sometimes preferred surgery as the initial treatment because they felt or were informed that surgery after radiotherapy failure could be more challenging due to fibrotic tissue.

> *"In the future, potential recurrent tumors cannot be treated or at least not as well [after radiotherapy]" patient 4*

The last medical factor of influence was the perceived center's experience with radiotherapy or surgery. Patients generally defined experience by the number of procedures yearly performed at the center. A number of patients directly enquired about the center's experience in VS surgery and radiotherapy and considered the perceived experience in their decision.

> *"They have the most experience. When you have to undergo such complex surgery, you want the best team there is" patient 13*

We identified several patient factors influencing the decision between surgery and radiotherapy: the patients' cognitions, the patients' experience, logistics and trust in the physician. In addition, anxiety played a role in the decision between surgery and radiotherapy. Anxiety about facial paresis specifically was reported by most patients.

> *"The neurosurgeon could not guarantee that no facial paresis would occur. In case that it would happen, it would be terrible." patient 1*

The anxiety about facial paresis did not differentiate between the two options because the perceived preservation of facial function was comparable. However, some patients favored surgery because they thought that surgical recovery of the facial nerve was possible only if the paresis was caused by surgery.

Other patients were afraid of surgical procedures inside the head and therefore favored radiotherapy. This anxiety was closely linked to other cognitions about the treatment. Some patients thought that radiotherapy was less invasive and, therefore, safer than surgery or, conversely, that radiotherapy did not solve anything. In addition, patients were influenced by their own or others' previous experiences with radiotherapy or surgery.

> *"I was thinking, opening my head and messing around, that was in my opinion not such a good idea" patient 3*

"because in my opinion, radiation does not solve anything" patient 6

Logistics was also a factor that influenced decision making. The perceived time investment associated with either surgery or radiotherapy and the perceived impact on work, study, normal daily life or holidays affected the decision between surgery and radiotherapy as well as the decision on timing of the start of treatment.

"The radiotherapy that is something I need to think about, every day traveling to the hospital and back is something I do not like" patient 12

Trust in physicians was the final factor that influenced decision making. Trusting the physician's capabilities and expertise was a prerequisite for choosing either surgery or radiotherapy. In addition, patients wanted to attain some level of affinity with their treating physician. A lack of trust, consisting of both confidence and affinity, made patients want a second opinion.

"They are the experts, but there should also be some connection, some human touch" spouse of patient 13

DISCUSSION

This qualitative study identified patient-reported factors that influence decisions in VS management. The decision making process entails one or two steps; the first step comprises the decision between active surveillance and active treatment. When active treatment is opted for, a second decision between the two active treatment modalities, radiotherapy or surgery, ensues. Both steps were influenced by factors that could be classified as medical or patient-related. Medical factors were tumor characteristics, physician's recommendation, treatment outcomes and center's experience with the treatment. Patient related factors were anxiety, personal characteristics, experience, cognitions, logistics and trust in physician.

Qualitative research enables researchers to find explanations for observations using the diversity of data and does not aim to provide generally transferable data. Although data saturation was reached, other themes may arise in different clinical or cultural contexts. Another challenge in qualitative research is to minimize the influence of the researcher's own preferences and assumptions. This is partly ensured by the use of the clear and transparent framework approach and the use of multiple coders, none of whom were directly involved in the patient care pathway.

Several medical factors have previously been described in quantitative research on VS. The physician's recommendation has been reported as the most influential factor in a patient's decision.[5,6,16,17] Tumor characteristics have been described as an additional influencing factor both in the decision between surgery and active surveillance and in the decision between treatment modalities.[5,16] Treatment outcomes such as change of symptoms and tumor control have also been reported as influencing factors in several studies, both quantitative and qualitative.[6,7,17] Our study added the theme of treatment failure, an important determinant.

Patient related factors on VS decision making that have been previously reported are anxiety and logistics.[7,8] An important aspect of the factor anxiety identified in our study is the perceived risk of facial paresis. This is in line with a study of Müller et al., reporting that patients ranked facial paresis as the most severe sequela.[6] Another aspect of anxiety is the perceived risk of complications of treatment. This was also reported in the qualitative study of Linkov et al.[7] The latter study also identified doubts about making the right decision as a factor, but this was not corroborated in the current study. Only one aspect of the factor logistics, i.e., return to work, has been previously identified in a decision trade-off study.[8]

Other patient-related factors identified in this study have not been previously reported for VS, but have been investigated in other diseases. For example, personal characteristics, such as decision making style and a patient's trust in the physician have been reported to influence treatment decisions in metastatic breast cancer.[18] The patient's cognitions about therapy and their own past experiences have been shown to influence management decisions in diabetes mellitus type II and lumbar disc herniation.[19,20] The patient's coping abilities and level of disease acceptance have been reported to affect treatment decisions in recurrent prostate cancer.[21]

Truly shared decision making requires the adequate and unbiased information of patients.[4] In addition to information provided by physicians, patients searched the internet for information about the disease, the treatment and experiences of other patients. The use of internet by otorhinolaryngology patients has increased over the years and has an increasing clinical impact.[22] However, the quality of online VS information varies highly.[23] It is important that physicians explore any preconceptions that a patient might have about the disease and the relevant treatment options in order to tailor the clinical information to the patients level of knowledge, deal with misconceptions if present and to ensure that patients fully understand the pros and cons of the treatment options.

Implications for practice

The findings of this study can be used to improve information and care provision in daily practice. In this study, one of the important medical factors that influenced the decision making was physician's recommendation. The physician's medical specialty will probably influence the provided recommendations, i.e., surgeons tend to advise surgery more often, whereas radiation oncologists tend to advise radiotherapy.[6,5] This could lead to unwanted practice variation. Moreover, the patients' cognitions about treatment, which were not always correct, also impacted the treatment decision. To overcome these problems, patients with small- to medium-sized VSs should be informed about all viable treatment options, preferably by all specialties involved in the different management strategies (radiotherapy, otorhinolaryngology and/or neurosurgery). This seems the best way to ensure balanced information on which patients base their decision.

In addition, the information provided during consultations could be better adapted to patient-related factors that influence decision making. Personal characteristics, cognitions and anxieties and their influence on the treatment decision can be addressed, but only if the physician is able to identify these factors.

Lastly, physicians should be aware that patients are also influenced by medically irrelevant factors such as accessibility of care, required time investment or even holiday planning.

Tailoring information provision to an individual patients' needs could enhance patient involvement in clinical decision making, which has been shown to reduce decisional conflict in VS patients.[24]

These insights into the factors that influence the patients' decision can be used to improve the decision making process in a number of ways. Firstly, patients with an indication for active treatment, in whom radiotherapy and surgery are both viable treatment options according to the MDT meeting, should be informed about both treatment modalities. To ensure balanced information about the effectiveness and downsides of radiotherapy and surgery, it is now provided by both a radiation oncologist and a surgeon (either a neurosurgeon or otorhinolaryngologist) in sequentially planned consultations at one of the participating centers. Secondly, patient-related factors could be better identified and monitored using patient reported outcome measures (PROMs) structurally. The factors that influence the patients'clinical decision as identified in this qualitative study can thus subsequently be evaluated in a quantitive way, in order to study their prevalence and relative importance. Anxiety, for example, is identified by the

anxiety subscale of a disease-specific quality of life questionnaire, the Penn Acoustic Neuroma Quality of Life (PANQOL).[25] In addition, to help patients to cope with their anxiety a psychologist has been added to the VS care team. Thirdly, public information on hospital websites and patient information flyers could be improved by involving patient representatives in order to better align the information with the patients' needs and expectations.

CONCLUSION

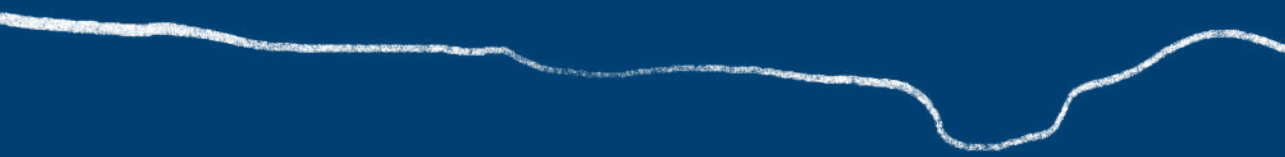
This study provides new insights into the factors that influence patients' decision making in small- to medium-sized VSs. Medical factors, such as tumor characteristics and the physician's recommendation were confirmed to play a role. In addition, new patient-related factors were identified, such as decision making style, the patients' trust in the physician, the patient's cognitions about therapy and past experiences and the patient's personal characteristics. Awareness of these factors is important for adequate patient counseling and may help in reaching truly shared VS management decisions.

REFERENCES

1. Carlson ML, Link MJ, Wanna GB, Driscoll CLW (2015) Management of Sporadic Vestibular Schwannoma. Otolaryngologic Clinics of North America, vol 48. doi:10.1016/j.otc.2015.02.003
2. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, Cording E, Tomson D, Dodd C, Rollnick S, Edwards A, Barry M (2012) Shared decision making: A model for clinical practice. Journal of General Internal Medicine, vol 27. Springer. doi:10.1007/s11606-012-2077-6
3. Stiggelbout AM, Van der Weijden T, De Wit MPT, Frosch D, Légaré F, Montori VM, Trevena L, Elwyn G (2012) Shared decision making: Really putting patients at the centre of healthcare. BMJ (Clinical research ed) 344:e256. doi:10.1136/bmj.e256
4. Stiggelbout AM, Pieterse AH, De Haes JCJM (2015) Shared decision making: Concepts, evidence, and practice. Patient Educ Couns 98:1172-1179. doi:10.1016/j.pec.2015.06.022
5. Moshtaghi O, Goshtasbi K, Sahyouni R, Lin HW, Djalilian HR (2018) Patient Decision Making in Vestibular Schwannoma: A Survey of the Acoustic Neuroma Association. Otolaryngology–Head and Neck Surgery 158 (5):912-916. doi:10.1177/0194599818756852
6. Müller S, Arnolds J, van Oosterhout A (2010) Decision-making of vestibular schwannoma patients. Acta Neurochir (Wien) 152:973-984. doi:10.1007/s00701-009-0590-0
7. Linkov F, Valappil B, McAfee J, Goughnour SL, Hildrew DM, McCall AA, Linkov I, Hirsch B, Snyderman C (2017) Development of an evidence-based decision pathway for vestibular schwannoma treatment options. Am J Otolaryngol 38 (1):57-64. doi:10.1016/j.amjoto.2016.09.019
8. Cheung SW, Aranda D, Driscoll CL, Parsa AT (2010) Mapping clinical outcomes expectations to treatment decisions: an application to vestibular schwannoma management. Otol Neurotol 31:284-293. doi:10.1097/MAO.0b013e3181cc06cb
9. Witt J, Elwyn G, Wood F, Brain K (2012) Decision making and coping in healthcare: The Coping in Deliberation (CODE) framework. Patient Educ Couns 88 (2):256-261. doi:10.1016/j.pec.2012.03.002
10. Flynn KE, Smith MA, Vanness D (2006) A typology of preferences for participation in healthcare decision making. Soc Sci Med 63 (5):1158-1169. doi:10.1016/j.socscimed.2006.03.030
11. Braun V, Clarke V (2006) Using thematic analysis in psychology. Qualitative Research in Psychology 3 (2):77-101. doi:10.1191/1478088706qp063oa
12. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA (2014) Standards for Reporting Qualitative Research: A Synthesis of Recommendations. Acad Med 89:1245-1251. doi:10.1097/ACM.0000000000000388
13. Pope C, Mays N (1995) Reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health services research. BMJ (Clinical research ed) 311:42-45
14. Smith J, Firth J (2011) Qualitative data analysis: the framework approach. Nurse Res 18:52-62. doi:10.1016/j.nedt.2010.12.011
15. Gale NK, Heath G, Cameron E, Rashid S, Redwood S (2013) Using the framework method for the analysis of qualitative data in multidisciplinary health research. BMC Med Res Methodol 13. doi:10.1186/1471-2288-13-117
16. Nellis JC, Sharon JD, Pross SE, Ishii LE, Ishii M, Dey JK, Francis HW (2017) Multifactor Influences of Shared Decision-Making in Acoustic Neuroma Treatment. Otol Neurotol 38:392-399. doi:10.1097/MAO.0000000000001292
17. Broomfield SJ, O'Donoghue GM (2016) Self-reported symptoms and patient experience: A British Acoustic Neuroma Association survey. Br J Neurosurg 30 (3):294-301. doi:10.3109/02688697.2015.1071323

18. Rocque GB, Rasool A, Williams BR, Wallace AS, Niranjan SJ, Halilova KI, Turkman YE, Ingram SA, Williams CP, Forero-Torres A, Smith T, Bhatia S, Knight SJ (2019) What Is Important When Making Treatment Decisions in Metastatic Breast Cancer? A Qualitative Analysis of Decision-Making in Patients and Oncologists. The Oncologist 24 (10):1313-1321. doi:10.1634/theoncologist.2018-0711
19. Lee YK, Low WY, Ng CJ (2013) Exploring Patient Values in Medical Decision Making: A Qualitative Study. PLoS One 8 (11):e80051. doi:10.1371/journal.pone.0080051
20. Andersen SB, Birkelund R, Andersen MØ, Carreon LY, Coulter A, Steffensen KD (2019) Factors Affecting Patient Decision-making on Surgery for Lumbar Disc Herniation. Spine (Phila Pa 1976) 44 (2):143-149. doi:10.1097/BRS.0000000000002763
21. Gorawara-Bhat R, O'Muircheartaigh S, Mohile S, Dale W (2017) Patients' perceptions and attitudes on recurrent prostate cancer and hormone therapy: Qualitative comparison between decision-aid and control groups. J Geriatr Oncol 8 (5):368-373. doi:10.1016/j.jgo.2017.05.006
22. Ihler F, Canis M (2019) Die Rolle des Internets für Gesundheitsinformationen in der Hals-Nasen-Ohrenheilkunde. Laryngo-Rhino-Otologie 98 (S 01):S290-S333. doi:10.1055/a-0801-2585
23. Spiers H, Amin N, Lakhani R, Martin AJ, Patel PM (2017) Assessing Readability and Reliability of Online Patient Information Regarding Vestibular Schwannoma. Otol Neurotol 38 (10):e470-e475. doi:10.1097/mao.0000000000001565
24. Graham ME, Westerberg BD, Lea J, Hong P, Walling S, Morris DP, Hebb ALO, Galleto R, Papsin E, Mulroy M, Foggin H, Bance M (2018) Shared decision making and decisional conflict in the Management of Vestibular Schwannoma: a prospective cohort study. Journal of Otolaryngology - Head & Neck Surgery 47 (1). doi:10.1186/s40463-018-0297-4
25. van Leeuwen BM, Herruer JM, Putter H, Jansen JC, van der Mey AG, Kaptein AA (2013) Validating the Penn Acoustic Neuroma Quality Of Life Scale in a sample of Dutch patients recently diagnosed with vestibular schwannoma. Otol Neurotol 34 (5):952-957. doi:10.1097/MAO.0b013e31828bb2bb

Data driven vestibular schwannoma care



Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (AI-PREM)

Marieke van Buchem
Olaf Neve
Ilse Kant
Ewout Steyerberg
Hileen Boosman
Erik Hensen

BMC Med Inform Decis Mak. 2022 Jul 15;22(1):183.
doi: 10.1186/s12911-022-01923-5.

ABSTRACT

Background

Evaluating patients' experiences is essential when incorporating the patients' perspective in improving healthcare. Experiences are mainly collected using closed-ended questions, although the value of open-ended questions is widely recognized. Natural language processing (NLP) can automate the analysis of open-ended questions for an efficient approach to patient-centeredness.

Methods

We developed the Artificial Intelligence Patient-Reported Experience Measures (AI-PREM) tool, consisting of a new, open-ended questionnaire, an NLP pipeline to analyze the answers using sentiment analysis and topic modeling, and a visualization to guide physicians through the results. The questionnaire and NLP pipeline were iteratively developed and validated in a clinical context.

Results

The final AI-PREM consisted of five open-ended questions about the provided information, personal approach, collaboration between healthcare professionals, organization of care, and other experiences. The AI-PREM was sent to 867 vestibular schwannoma patients, 534 of which responded. The sentiment analysis model attained an F1 score of 0.97 for positive texts and 0.63 for negative texts. There was a 90% overlap between automatically and manually extracted topics. The visualization was hierarchically structured into three stages: the sentiment per question, the topics per sentiment and question, and the original patient responses per topic.

Conclusions

The AI-PREM tool is a comprehensive method that combines a validated, open-ended questionnaire with a well-performing NLP pipeline and visualization. Thematically organizing and quantifying patient feedback reduces the time invested by healthcare professionals to evaluate and prioritize patient experiences without being confined to the limited answer options of closed-ended questions.

BACKGROUND

Patient-centeredness is an essential fundament for providing high-quality care.[1, 2] Insight into the patient-centeredness of care is obtained by evaluating patient experiences, typically using Patient-Reported Experience Measures (PREMs). Most PREMs include a combination of closed- and open-ended questions. When presented with both, healthcare professionals tend to value the answers to open-ended questions most [3]. These answers can be used to identify new points of interest ('topics') and provide context to closed-ended questions.[3, 4] Although the value of open-ended questions is widely recognized, patients' free-text answers remain underutilized in clinical practice. One of the key challenges lies in the time needed for analysis. The answers to open-ended questions are often manually analyzed, which is laborious and time-consuming[3], especially in larger groups of patients.

Several studies aim to automate the analysis of free-text patient experience data to inform quality improvements, showing promising results.[5-15] Most of these studies concentrate on publicly available social media or forum data, usually focused on reviewing hospitals or physicians.[5-9] Current approaches include the use of artificial intelligence (AI) methods such as machine learning and natural language processing (NLP). A few studies successfully applied NLP techniques to routinely collected PREM questionnaires of patients.[10-15] Most of these studies use supervised methods; for example, topic classification is used to classify data into predefined, manually extracted topics.[5, 7, 11, 13, 15] Although some of these methods perform well, supervised methods lack the capability of finding new or unexpected topics. Moreover, regular manual labeling is time-consuming and, therefore, not suited to decrease the current burden of reading through the patients' answers.[10] Using unsupervised methods such as topic modeling can overcome these limitations. Two studies have compared supervised topic classification to unsupervised topic modeling and concluded that topic modeling leads to topics similar in quality.[7, 15]

Current open-ended questions are often unsuitable for automatic analysis as they were not developed for this purpose.[10, 11] An example is a questionnaire consisting of the questions 'What did we do well?' and 'What could we improve?'. Previous work shows that answers to both questions can be positive and negative, complicating automated sentiment analysis.[10, 11] One study created a new, open-ended questionnaire suitable for analysis with NLP[16], focusing on patient-reported outcomes instead of experiences. They concluded that adding open-ended questions leads to richer, more in-depth information, and analysis with NLP makes it feasible to use in clinical practice.

The aim of this study is to harness the value of free-text patient experiences, using NLP methods that have the flexibility to find new topics in a complex, fast-changing environment. Our approach is to develop and validate a method for collecting and analyzing open-ended PREMs that could be incorporated into clinical practice. This objective contains three sub-objectives:

1. Develop and validate an open-ended generic PREM questionnaire;
2. Develop and validate an NLP pipeline to automatically analyze the open-ended PREM;
3. Develop a visualization that supports healthcare professionals in identifying quality improvements from the results.

METHODS

We devised a method that included a new, open-ended questionnaire, an NLP pipeline to analyze the questionnaire, and a visualization of the output of the NLP pipeline (Fig. 1). This project was organized in a development phase and a validation phase. The development phase started with developing a new questionnaire, the Artificial Intelligence Patient-Reported Experience Measure (AI-PREM).

Development of the AI-PREM (Fig. 1, step 1)

The AI-PREM was developed iteratively with patients from the vestibular schwannoma care pathway in the Leiden University Medical Center (LUMC) (Box 1). We used the following criteria: (1) Open-ended questions; (2) Phrasing suitable for analysis with NLP;

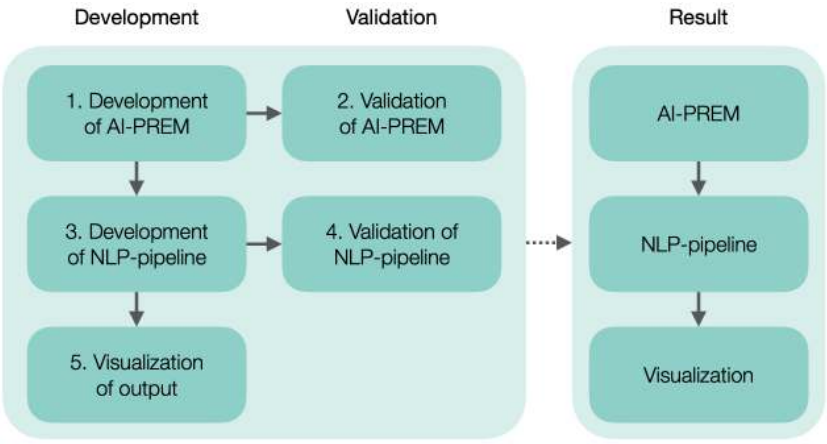


Fig. 1 Overview of the different tasks and phases

Box 1 Description of the vestibular schwannoma care pathway in the LUMC

Vestibular schwannomas are benign intracranial tumors, with a heterogeneous clinical presentation: it may present as a small, slow growing, and asymptomatic tumor, but also as large, faster growing, and potentially fatal disease. Patients typically present with symptoms of hearing loss, loss of balance and vertigo, but may also suffer from facial numbness, facial paralysis, or elevated intracranial pressure. In non-progressive tumors, active surveillance with MRI is usually the management option of choice. In progressive tumors, surgery or radiotherapy is performed to prevent future complications. After an active intervention, prolonged active surveillance ensues in these patients too, in order to identify possible recurrences. The LUMC is an expert referral center for vestibular schwannoma in the Netherlands. The care is organized in an integrated practice unit including all specialties involved in the diagnosis and treatment (i.e., neurosurgery, otorhinolaryngology, radiology and radiation oncology).

(3) Generic questions, therefore not containing disease-, department-, or center-specific questions; (4) Accessible in terms of length and language. The Picker principles of patient-centered care[17] were the basis for the questionnaire. The development process started with questions about all eight Picker principles, asking patients about experiences with the accessibility of care, continuity of care, involvement of family, emotional support, information provision, physical needs, and involvement in decisions. Each question included one subject and did not contain a sentiment, to decrease the variability of patients' answers. For example, instead of asking 'What could be improved in the organization of care?' the question stated 'How was the organization of care?'. These questions were evaluated and finetuned in a group of patients.

Patients who participated in a survey study in 2014 were re-approached for participation in the AI-PREM project between May and September 2020.[18] Patients that agreed to participate provided their written informed consent. All patients were diagnosed with unilateral vestibular schwannoma between 2003 and 2014. Patients with bilateral vestibular schwannoma, other skull base pathologies, or insufficient proficiency in the Dutch language to complete the questionnaires were excluded. In addition to the AI-PREM, patients were also asked to fill out a validated structured patient experience questionnaire, the patient experience monitor (PEM), for comparison.[1] Patients first filled out the AI-PREM to ensure they were not biased towards the topics assessed in the PEM. The questionnaires were sent out either by e-mail using Castor software or hard copy by mail. These hard copies were verbatim digitalized manually.

Validation of the AI-PREM (Fig. 1, step 2)

To validate the AI-PREM questionnaire, we used the COSMIN reporting guideline for studies on the measurement properties of patient-reported outcome measures.[19] Although this guideline is aimed at structured questionnaires about patient outcomes, most parts can be applied to unstructured patient experience questionnaires. The COSMIN guideline investigates the content validity of questionnaires by looking at the questions' relevance, comprehensiveness, and comprehensibility. We examined the content validity of the AI-PREM by comparing AI-PREM questions to similar questions from the PEM. First, a sentiment analysis (as described in the Sentiment analysis section under 'Development of the NLP pipeline') was performed, labeling a text as positive, neutral or negative feedback. We hypothesized that patients who were negative about certain aspects of care in the AI-PREM would also give lower scores on the matched PEM questions and vice versa (scores range from one to ten, where one is the lowest and ten is the highest). Therefore, we defined 'positive' and 'negative' comments per AI-PREM question based on the sentiment analysis. Per AI-PREM question, we took the matched PEM questions and calculated the average score for the 'positive' and 'negative' groups. Using a t-test for independent samples, we compared the average scores between the 'positive' and 'negative' groups.

Development of the NLP pipeline (Fig. 1, step 3)

The pipeline as described by Cammel et al. was taken as a starting point.[10] The pipeline includes sentiment analysis, preprocessing, and topic modeling. We combine a supervised (sentiment analysis) and unsupervised (topic modeling) approach. We use a supervised approach for the sentiment analysis because the categories for this task (positive, neutral, negative) will not change over time, in contrast to the topics that patients mention. The pipeline was developed in an iterative process by a team of data scientists, researchers, and clinicians of the vestibular schwannoma IPU, to fulfill the following pre-set requirements:

- Interpretable: The end-user should be able to distill from the output what patients experience as positive and negative.
- Actionable: The output should be specific enough to lead to concrete action points.
- Complete: The number of texts that cannot be assigned to a topic should be as small as possible.

Once the output met all the requirements according to the development team, the validation phase started.

Sentiment analysis

We finetuned a pretrained, multilingual BERT model for two binary classification tasks for sentiment analysis. The first binary classification task classified answers as negative or non-negative; the second task classified the non-negative answers as positive or neutral. To train these two sentiment analysis models, one annotator (MvB) manually labeled 75% of the collected data as 'negative', 'positive', or 'neutral'. A second annotator (ON) labeled 1/3rd of this data (25% of the collected data), which was used to calculate the inter-annotator agreement (percentage of datapoints that the annotators agreed on). Annotators labeled an answer as 'negative' if it described a topic or situation that the patient was dissatisfied with (e.g., 'I had to wait for a long time'). If a non-negative answer described a topic or situation that the patient was satisfied with, it was labeled as 'positive' (e.g., 'the personnel was very friendly'). All answers that described a topic or situation that was neither positive nor negative were labeled as 'neutral' (e.g., 'first I was treated at hospital number 1, then I was referred to hospital number 2'). The two sentiment analysis models were trained on a random sample of 80% and validated on the other 20% of labeled data, using the default parameters of the Transformers implementation of the BERT model for Sequence Classification.[20]

Preprocessing

After the sentiment analysis, the data were preprocessed. We tokenized words and corrected the spelling using the Peter Norvig algorithm[21] and the CyHunSpell Python package[22]. Subsequently, words were lemmatized, and all non-informative words (stopwords, words with less than three letters, and all words except verbs, adverbs, nouns, and adjectives) were removed using the Stanza Python package[23]. Finally, all n-grams ranging from one to three were vectorized using term frequency-inverse document frequency (TF-IDF).

Topic modeling

We used topic modeling, specifically Non-negative Matrix Factorization (NMF), to identify the most important topics from the patients' answers to the AI-PREM, as described by Cammel et al.[10] NMF was chosen over Latent Dirichlet Allocation because patients' answers tend to be very short and NMF is better able to deal with short answers. A separate topic model was created per sentiment (positive or negative) and per question. For each topic model, the optimal number of topics was chosen by creating several topic models with topics ranging from 2 to 15 and calculating the coherence score within every topic. The coherence score was calculated using the semantic similarity of words within a topic, based on a Dutch Word2Vec model[24,25,26], to account for exact matches and synonymous words. The topic model with the highest coherence metric was chosen as the best fitting model for that specific category.

Validation of the NLP pipeline (Fig. 1, step 4)

We performed different validation steps to evaluate the performance of the NLP pipeline. (1) We assessed whether the automatically defined topics were representative of the texts they described. (2) We evaluated whether the NLP pipeline extracted topics similar to human-extracted topics.

Representativeness of topics

We randomly sampled the answers to the AI-PREM and performed manual evaluations of these answers by clinical experts. One clinician (ON) assessed a sample of the texts within the different categories (e.g., positive answers about information, negative answers about the organization of care). Per category, 20% of the answers per topic were analyzed, with a minimum of 10 texts. Some topics included less than ten texts; the clinician evaluated all texts for these topics. For every text within the sample, the clinician decided if it fit within the assigned topic. This analysis resulted in a percentage showing how representative the different topics were for the answers within that topic. A researcher (MvB) went through the same validation process to calculate the inter-annotator agreement.

Topic model versus human comparison

To investigate the performance of the topic model compared to human analysis, two clinical experts (a physician and a nurse practitioner) from the vestibular schwannoma care pathway read the answers to the AI-PREM from a sample of 50 patients, as data saturation was reached. A qualitative approach was used to identify topics within these texts. After reading, the experts decided on a few topics per question that summarized patients' answers in a consensus meeting. Two researchers (MvB and ON) compared these manually selected topics to the automatically selected topics from the NLP pipeline. Because the human analysis consisted of a sample of 50 questionnaires (and not all), we did not try to match exact words but matched on topic level. The proportion of manually identified topics that could be matched to an automatically identified topic was subsequently calculated.

Visualization of the output (Fig. 1, step 5)

To stimulate the use of the AI-PREM tool in clinical practice, we co-created a mock-up of a potential visualization. We held three feedback sessions with a group of physicians, nurse practitioners, and implementation managers and iteratively updated the visualization based on their feedback and pre-set requirements. The requirements for the visualization were:

1. Applicability within the end-users current workflow;
2. Presentation of an overview of the output at a glance;
3. Ability to get more context without going through all the individual questionnaires.

RESULTS

Development of the AI-PREM

During six iterations, the initial questions were finetuned. The most significant changes made during these iterations were reducing the number of questions and simplifying the sometimes abstract Picker principles. The comprehensibility improved by using only level B1 words of the Common European Framework of Reference for Languages.[27] Furthermore, patients preferred to have some examples of what was meant by the different aspects. The Picker institute provides some examples, which we added to each question. This led to the following questions:

- Q1: How was the provided information? Think of: the prognosis, possible tests, and treatment(s)
- Q2: How was the personal approach? Think of: shared decision making, listening to your preferences, emotional support
- Q3: How was the collaboration between healthcare professionals? Think of: no varying advice or having to tell your story multiple times, contact with your family doctor or other hospitals
- Q4: How was the organization of care? Think of: making appointments, combining appointments on one day, availability by phone
- Q5: What else would you like to share about your experience?

In total, 536 out of 867 vestibular schwannoma patients filled out the AI-PREM and PEM questionnaires, resulting in a response rate of 62%. Two patients were excluded because their diagnosis changed from vestibular schwannoma to meningioma, requiring treatment in another care pathway. This resulted in 534 sets of questionnaires. The median length of patients' answers was two words, with an interquartile range of 1 to 11 words. The maximum length was 192 words.

Validation of the AI-PREM

Using the Picker principles as a basis, the AI-PREM adhered to the relevance and comprehensibility criteria from the COSMIN reporting guideline. The comprehensibility criterium was further substantiated by including patients in the development of the AI-PREM. The results of validating the last criterium, comprehensiveness, are shown in Table 1. Where Q1-3 showed a significant difference in PEM scores between positive and negative answers, Q4 did not. No PEM questions were matched to Q5 ('What else would you like to share about your experience?'), so we did not validate this question.

Table 1 Overview of the number of AI-PREM responses per sentiment

Questions	Number of patients N (%)	Average PEM scores of matched questions, ranging from 1 to 10 μ ± sd
Q1–Positive	359 (67.2%)	9.7 ± 0.9
Negative	26 (4.9%)	8.1 ± 2.4**
Q2–Positive	360 (67.4%)	9.7 ± 0.7
Negative	31 (5.8%)	7.7 ± 2.6**
Q3–Positive	325 (60.9%)	9.6 ± 1.1
Negative	40 (7.5%)	8.3 ± 1.8*
Q4–Positive	343 (64.2%)	6.9 ± 1.7
Negative	39 (7.3%)	6.4 ± 2.0
Q5–Positive	121 (22.7%)	
Negative	35 (6.6%)	

The neutral responses are left out. Per category (question and sentiment), the average scores to the PEM questions that matched the AI-PREM questions are shown. P-value for the t-test for independent samples: * = $p < 0.001$, ** = $p < 0.0001$. AI-PREM: artificial intelligence patient reported experience measure. PEM: patient experience monitor. Q: question. sd: standard deviation

Development of the NLP pipeline

We made several improvements to the pipeline during the iterative development process (Box 2). The final NLP pipeline contained a sentiment analysis model consisting of a negative and positive sentiment classifier and a topic modeling module (Fig. 2).

Box 2 Most important improvements that were made during the iterative development process

- To first perform a sentiment analysis and then create a separate topic model per sentiment and per question, instead of creating one topic model for both sentiments. This led to more specific topics, from which points of improvements could be derived more easily, increasing the interpretability and actionability
- To not only include the negative feedback topics but also the positive ones, in order to obtain more balanced information. This was found to be essential in selecting and prioritizing points of improvement. In addition, the positive topics were seen as motivators for the healthcare team
- To go from a fixed number of topics to an adaptive approach that automatically chooses the optimal number of topics per subject. This increased the completeness
- To add a quantitative dimension to the qualitative output of the topic model, in order to help prioritize aspects of care that need the most attention
- To include n-grams up to three instead of just using 1 g. This increased the interpretability and actionability of the topics

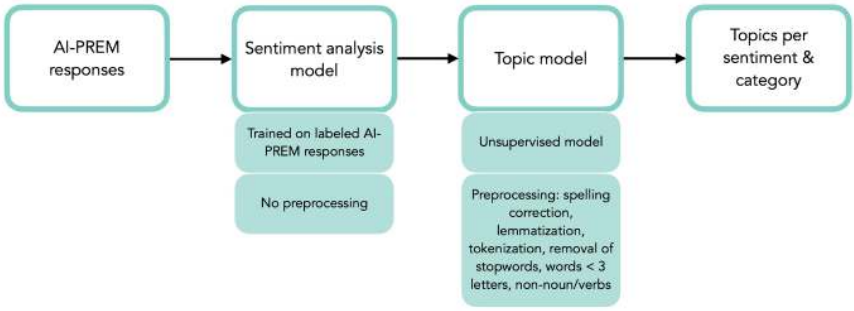


Fig. 2 Overview of the input, models, and output of the AI-PREM tool

Sentiment analysis

The inter-annotator agreement was 91.9%. The precision and recall for the negative sentiment model were 0.78 and 0.53, respectively, with an F1 score of 0.63. The precision, recall, and F1 score for the positive sentiment model were all 0.97.

Topic modeling

The number of topics per category ranged from two to six. 2.8% of texts could not be assigned to a topic. Only the ten n-grams with the highest TF-IDF score per topic were extracted to increase the interpretability of the topics. These n-grams were sorted based on the number of words, with the highest number of words shown first. We deduplicated this list of words to ensure that the final list of descriptors would not contain both 'went very well' and 'went well'. Finally, the first five words of this sorted, deduplicated list were shown to the end-user (Fig. 3). See Additional file 1 for all the topics per category.

Q5: What else would you like to share about your experience?

Positive topics	Amount
Only good, good experience, experience good, aftercare good, very good	105
Treatment aftercare, only positive, only good, good experience, everyhing fine	8

Negative topics	Amount
Aftercare good, aftercare deal with, deal with new, situation well, new situation	14
Long wait, result scan, wait result, long ago, surgery confess	21

Fig. 3 Topic model for Q5

Validation of the NLP pipeline

The overall percentage of representative texts was 80.9%, with 90.1% for the positive texts and 72.0% for the negative texts (Table 2). The inter-annotator agreement was 94.4% for positive texts, 80.5% for negative ones, and 90.4% overall. The clinical experts extracted 20 topics: 14 for the positive and 6 for the negative texts. All negative topics

Table 2 Representativeness of the different topic models per category

Question	Positive categories in total	Per topic	Negative categories in total	Per topic
Q1	94.4% (n = 72)	T1: 100% (n = 36) T2: 88.9% (n = 36)	55.6% (n = 18)	T1: 60% (n = 10) T2: 50% (n = 8)
Q2	93.3% (n = 75)	T1: 97.1% (n = 35) T2: 100% (n = 10) T3: 85% (n = 20) T4: 90% (n = 10)	71% (n = 31)	T1: 100% (n = 3) T2: 100% (n = 3) T3: 83.3% (n = 6) T4: 100% (n = 3) T5: 75% (n = 4) T6: 28.6% (n = 7) T7: 60% (n = 5)
Q3	98.4% (n = 63)	T1: 100% (n = 43) T2: 95% (n = 20)	76.9% (n = 39)	T1: 100% (n = 4) T2: 33.3% (n = 3) T3: 85.7% (n = 7) T4: 100% (n = 5) T5: 66.7% (n = 3) T6: 77.8% (n = 9) T7: 62.5% (n = 8)
Q4	100% (n = 65)	T1: 100% (n = 41) T2: 100% (n = 12) T3: 100% (n = 12)	86.7% (n = 15)	T1: 100% (n = 5) T2: 80% (n = 10)
Q5	86.2% (n = 29)	T1: 85.7% (n = 21) T2: 87.5% (n = 8)	55.5% (n = 20)	T1: 50% (n = 10) T2: 60% (n = 10)

Representativeness is defined as the number of texts within a certain topic that fit the description of the topic. The percentage is calculated by dividing the texts that fit the description of the topic by the total number of texts within the topic.
Q: AI-PREM question. T: automatically extracted topic

and 12 of 14 positive topics could be matched to the automatically extracted topics, leading to a 90% overlap between human topics and automatically extracted topics.

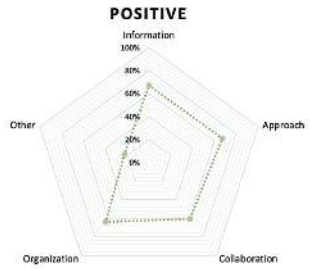
Visualization of the output

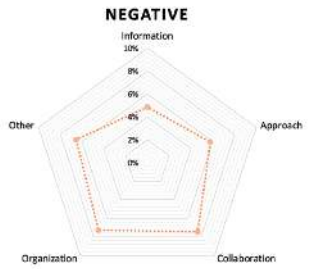
The end-users preferred the spider plot over other visualizations in the feedback session, such as a bar plot or tornado graph. The final visualization included a mock-up with three stages (Fig. 4).

DISCUSSION

This study describes the development and validation of a comprehensive tool for surveying the patient experience that can automatically produce actionable information. The tool consists of an open-ended, validated patient experience questionnaire suitable for qualitative and quantitative analysis with natural language processing (NLP), a well-performing NLP pipeline to analyze the answers to the questionnaire automatically, and

(a)





What else would you like to share about your experiences?

Positive clusters	Amount
Only good, good experience, experience good, aftercare good, very good	105
Treatment aftercare, only positive, only good, good experience, everything fine	8

Negative clusters	Amount
Aftercare good, aftercare deal with, deal with new, situation well, new situation	14
Long wait, result scan, wait result, long ago, surgery confess	21

(b)

What else would you like to share about your experiences?

Positive topic 1	Amount
Only good, good experience, experience good, aftercare good, very good	105

Negative topic 1	Amount
Aftercare good, aftercare deal with, deal with new, situation well, new situation	14

Original patient responses

The interaction with all the staff was very good, they really listened to me
The contact with the nurse and doctor was very good
Good support by the nurse
Good
Good collaboration between the other hospital and the LUMC. It was easy to book a new scanning appointment and the information was clear.
I can only be positive about the treatment and the aftercare.

Aftercare could be improved
Aftercare could be slightly better
Exceptionally difficult to make the right decision. Thoroughly delved into the problem. Eventually chose the direct approach in the form of Cyberknife radiation.
Aftercare (dealing with new situation) could be better
Aftercare female eye doctor was so-so

Fig. 4

a Stage 1: the spider plot showing the percentage of positive and negative texts per question. Stage 2: once the end-user clicks on one of the questions, the automatically extracted topics are shown. The positive topics are shown on the left and the negative topics on the right.

b Stage 3: if the end-user wants to dive into one of the topics, they can click on that topic and read the actual patient answers that belong to that topic. In this example, the end-user is looking at the topics within the 'Other' category and has clicked on positive topic 1 and negative topic 1

a visualization that supports healthcare professionals in defining quality improvements from the results.

A critical aspect of our study is that we created and validated a new questionnaire consisting of only open-ended questions. One other study developed a new, open-ended questionnaire suitable for analysis with NLP, but they focused on patient outcomes

instead of experiences.[16] Unique in our study is that we compared the AI-PREM with a 'gold standard' PREM, the patient experience monitor (PEM). Overall, three out of four open-ended questions of the AI-PREM seem to capture sentiments similar to the PEM. The lack of a significant correlation for the fourth question, asking about the organization of care, might be explained because this question had the lowest average PEM score and the smallest range.

Our NLP pipeline combines sentiment analysis with topic modeling while also making it possible to go back to individual patients' original responses per topic. This hierarchical structure allows healthcare professionals to scan the sentiment analysis for a high-level view or dive into the different topics and texts to define quality improvements. Physicians can use the quantitative data to review the results at a glance and prioritize the various topics, while the qualitative data allows them to put the topics into context and define concrete points of action.

Unlike most studies[5, 7, 11, 13, 15], we chose an unsupervised topic modeling approach due to its flexibility in finding new and unexpected topics.[3, 10] One example that highlights the benefit of this approach is the topic describing the negative sentiment patients had about how long they had to wait for the scan results. This topic is not included in structured questionnaires and is very specific to this care pathway. Furthermore, the differing number of topics per question shows the ability of this method to adapt to the data at hand. Methods sensitive to changing topics in patients' experiences are essential in the constantly changing healthcare environment.

We finetuned a pretrained multilingual BERT model on our data for the current sentiment analysis. Because the questionnaire and answers were in the Dutch language, there was limited choice in off-the-shelf sentiment analysis models, and the available models did not perform well on our data. Furthermore, there are no BERT models pretrained on clinical data for Dutch, so we used the multilingual BERT model as a basis. The positive sentiment model performs better than most other studies, with an F1 score of 0.97. Other studies report F1 scores between 0.74 and 0.90 for sentiment analysis on patient experience data.[6, 14, 15, 28, 29] The negative sentiment model performs below average, with an F1 score of 0.63. The small number of negative texts compared to the amount of neutral and positive texts causes this difference. With more data, the model can be trained further to improve the performance in recognizing negative texts and make it more generalizable to other departments and care pathways.

Our manual validation of the NLP pipeline shows that the quality of the topics is high in terms of the representativeness of the topics and the similarity to the manual topics.

These results align with previous studies that show the similarity between supervised, manually defined topics and unsupervised, automatically defined topics.[7, 15] However, there is a large difference in the quality of the topics for the different categories in the AI-PREM. Although most topics represent their texts very well with scores ranging from 90 to 100%, a few mostly negative topics have scores between 20 and 50%. One possible explanation is the heterogeneity in the negative answers, leading to a few 'left-over' topics that fail to represent the texts well. One solution would be to gather more data before running the model, as this would decrease the chance of getting topics that only contain a few texts. Another solution is changing the phrasing of the questionnaire by making it more specific or giving different examples. Especially the question about the organization could be improved because this question also showed low responsiveness to changes in sentiment. On the other hand, the number of texts that could not be assigned a topic was only 2.8%, which is much better than the 15.4% reported in previous work.[10] It shows that a larger amount of texts can be automatically analyzed and confirms the improved suitability of our proposed open-ended questions for NLP analysis. In a previous report by Spasíc et al.[16], the authors optimize their questionnaire comprising open-ended questions in a similar way, i.e., by focusing every question on one particular aspect (different patient outcomes in their case), extracting any sentiment from the question itself, and providing examples per question (also at their patients' request).

We noted that positive comments are much more numerous, but negative topics tend to be more elaborately discussed by patients. For example, the negative topics' wait result scan' and 'contact (with) other hospital' contain concrete problems, while 'information good' and 'only positive' are much more high-level. These results align with other studies [3, 11, 30], which also found more specific feedback in negative comments. As we aimed to facilitate the quality improvement process, we see no limitation in this finding: the in-depth nature of the negative feedback makes it possible to define specific points of improvement, while the more general positive feedback functions as motivation for healthcare professionals. Moreover, previous work on structured patient experience questionnaires describes the problem of the ceiling effect: patient experience questionnaires tend to overestimate patient satisfaction[4], and very satisfied patients often still include a point of improvement.[5, 31] The AI-PREM shows this same trend towards positive responses, but the ability to provide a free text response leads to more in-depth feedback. The tool further facilitates healthcare professionals to put topics into perspective by comparing positive to negative topics and forming concrete action points by going back to patients' original responses.

Strengths & limitations

A strength is the combination of quantitative data from the sentiment analysis and qualitative data from the topic models, which creates a clear, usable overview of patients' experiences. It also aligns with the proposed framework for automated analysis of opinionated data from a recent study.[32] This framework presents a similar pipeline, with sentiment analysis for the quantitative analysis followed by a more qualitative approach using, for example, topic modeling.

Another strength of the current study is the validation steps we took to assess the performance of the AI-PREM tool. Although it was challenging to find suitable validation methods, the current methods combined with the COSMIN reporting guideline provide some insight into how well the topics represent the patients' answers. However, the combination of the small sample size per topic and lack of easily interpretable metrics limits the use of topic modeling. Therefore, we could not compare our topic models to other literature.

The current sentiment analysis model, which assigns a whole text as either 'positive', 'neutral', or 'negative', is limited. By assigning texts as 'negative' if they contained at least one aspect that the patient was negative about, we made sure not to miss any points for improvement. However, in the future, we would like to finetune the model to define a sentiment per sentence instead of per text and to change the sentiment into a 5-point scale ranging from 'very dissatisfied' to 'very satisfied'. This granularity would make it easier to define priorities based on the level of dissatisfaction with a specific aspect of care.

Lastly, our current tool was built and validated in close consultation with clinicians, which ensures the internal validity of the model and clinically relevant and actionable output. However, it was validated using the patient experiences of a specific patient group. To investigate the generalizability of the AI-PREM tool, we will have to collect AI-PREM data in other patient groups and evaluate its usability for different groups of physicians.

CONCLUSIONS

The AI-PREM tool is a comprehensive method that combines a validated questionnaire consisting of open-ended questions with a well-performing NLP pipeline and visualization. By thematically organizing and quantifying patient feedback, it reduces the time

invested by healthcare professionals to evaluate and prioritize patient experiences without being confined to the limited answer options of closed-ended questions.

REFERENCES

1. Bastemeijer CM, Boosman H, Zandbelt L, Timman R, de Boer D, Hazelzet JA. Patient experience monitor (PEM): the development of new short-form picker experience Questionnaires for hospital patients with a wide range of literacy levels</p>. Patient Relat Outcome Meas. 2020;11:221–30.
2. Medicine I of. Crossing the Quality Chasm: A New Health System for the 21st Century. 2001; Available from: https://www.nap.edu/catalog/10027/crossing-the-quality-chasm-a-new-health-system-for-the
3. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. Bmj Heal Care Inform. 2021;28(1): e100262.
4. Riiskjaer E, Ammentorp J, Kofoed PE. The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective. Int J Qual Health C. 2012;24(5):509–16.
5. Alemi F, Torii M, Clementz L, Aron DC. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. Qual Manag Health Ca. 2012;21(1):9–19.
6. Anjum A, Zhao X, Bahja M, Lycett M. Identifying patient experience from online resources via sentiment analysis and topic modelling. Proc 3rd Ieee Acm Int Conf Big Data Comput Appl Technologies. 2016;94–9.
7. Jones J, Pradhan M, Hosseini M, Kulanthaivel A, Hosseini M. Novel approach to cluster patient-generated data into actionable topics: case study of a web-based breast cancer forum. JMIR Med Inform. 2018;6(4): e45.
8. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Machine learning and sentiment analysis of unstructured free-text information about patient experience online. Lancet. 2012;380:S10.
9. Ranard BL, Werner RM, Antanavicius T, Schwartz HA, Smith RJ, Meisel ZF, et al. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. Health Affair. 2017;35(4):697–705.
10. Cammel SA, Vos MSD, van Soest D, Hettne KM, Boer F, Steyerberg EW, et al. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. Bmc Med Inform Decis. 2020;20(1):97.
11. Khanbhai M, Warren L, Symons J, Flott K, Harrison-White S, Manton D, et al. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. Int J Med Inform. 2022;157: 104642.
12. Menendez ME, Shaker J, Lawler SM, Ring D, Jawa A. Negative patient-experience comments after total shoulder arthroplasty. J Bone Joint Surg. 2019;101(4):330–7.
13. Rivas C, Tkacz D, Antao L, Mentzakis E, Gordon M, Anstee S, et al. Automated analysis of free-text comments and dashboard representations in patient experience surveys: a multimethod co-design study. Heal Serv Deliv Res. 2019;7(23):1–160.
14. Nawab K, Ramsey G, Schreiber R. Natural language processing to extract meaningful information from patient experience feedback. Appl Clin Inform. 2020;11(02):242–52.
15. Doing-Harris K, Mowery DL, Daniels C, Chapman WW, Conway M. Understanding patient satisfaction with received healthcare services: A natural language processing approach. In: AMIA annual symposium proceedings. 2017.
16. Spasić I, Owen D, Smith A, Button K. KLOSURE: closing in on open–ended patient questionnaires with text mining. J Biomed Semant. 2019;10(Suppl 1):24.

17. Davis K, Schoenbaum SC, Audet AM. A 2020 vision of patient-centered primary care. J Gen Intern Med. 2005;20(10):953–7.
18. Soulier G, van Leeuwen BM, Putter H, Jansen JC, Malessy MJA, van Benthem PPG, et al. Quality of life in 807 patients with vestibular schwannoma: comparing treatment modalities. Otolaryngology Head Neck Surg. 2017;157(1):92–8.
19. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. Qual Life Res. 2021;30(8):2197–218.
20. Face H. BERT [Internet]. [cited 2021 Dec 14]. Available from: https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification
21. Norvig P. How to Write a Spelling Corrector [Internet]. 2016 [cited 2021 Nov 21]. Available from: https://norvig.com/spell-correct.html
22. Seal M, Rodriguez T. CyHunSpell [Internet]. 2021. Available from: https://pypi.org/project/cyhunspell/
23. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations [Internet]. 2020. Available from: https://nlp.stanford.edu/pubs/qi2020stanza.pdf
24. Tulkens S, Emmery C, Daelemans W. Evaluating unsupervised dutch word embeddings as a linguistic resource. In: Proceedings of the tenth international conference on language resources and evaluation (LREC 2016). European language resources association (ELRA); 2016.
25. Schäfer R, Bildhauer F. Building Large Corpora from the Web Using a New Efficient Tool Chain. Piperidis"] ["Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Asuncion Moreno and Jan Odijk and Stelios, editor. 23AD;486–93. Available from: http://rolandschaefer.net/?p=70
26. Schäfer R. Processing and querying large web corpora with the COW14 architecture. Witt"] ["Piotr Bański and Hanno Biber and Evelyn Breiteneder and Marc Kupietz and Harald Lüngen and Andreas, editor. 2015; Available from: http://rolandschaefer.net/?p=749
27. Europe C of. Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume [Internet]. Strasbourg: Council of Europe Publishing; 2020. Available from: www.coe.int/lang-cefr
28. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform. 2020. https://doi.org/10.2196/17984.
29. Jiménez-Zafra SM, Martín-Valdivia MT, Maks I, Izquierdo R. Analysis of patient satisfaction in Dutch and Spanish online reviews. Procesamiento del Lenguaje Natural. 2017;58:101–8.
30. Wagland R, Recio-Saucedo A, Simon M, Bracher M, Hunt K, Foster C, et al. Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care. Bmj Qual Saf. 2016;25(8):604.
31. Gallan AS, Girju M, Girju R. Perfect ratings with negative comments: learning from contradictory patient survey responses. Patient Exp J. 2017;4(3):15–28.
32. Kazmaier J, van Vuuren JH. A generic framework for sentiment analysis: leveraging opinion-bearing data to inform decision making. Decis Support Syst. 2020;135: 113304.

The added value of the artificial intelligence patient-reported experience measure (AI-PREM) tool in clinical practice: Deployment in a vestibular schwannoma care pathway

Olaf Neve
Marieke van Buchem
Marleen Kunneman
Peter Paul van Benthem
Hileen Boosman
Erik Hensen

PEC Innov. 2023 Aug 30:3:100204.
doi: 10.1016/j.pecinn.2023.100204

ABSTRACT

Objectives

Patient-reported experience measures (PREMs) can be used for the improvement of quality of care. In this study, the outcome of an open-ended question PREM combined with computer-assisted analysis is compared to the outcome of a closed-ended PREM questionnaire

Methods

This survey study assessed the outcome of the open-ended questionnaire PREM and a close-ended question PREM of patients with unilateral vestibular schwannoma in a tertiary vestibular schwannoma expert centre

Results

The open-ended questions PREM, consisting of five questions, was completed by 507 participants and resulted in 1508 positive and 171 negative comments, categorised into 27 clusters. The close-ended questions PREM results were mainly positive (overall experience graded as 8/10), but did not identify specific action points. Patients who gave high overall scores (>8) on the close-ended question provided points for improvement in the open-ended question PREM, which would have been missed using the close-ended questions only.

Conclusions

Compared to the close-ended question PREM, the open-ended question PREM provides more detailed and specific information about the patient experience in the vestibular schwannoma care pathway.

Innovation

Automated analysis of feedback with the open-ended question PREM revealed relevant insights and identified topics for targeted quality improvement, whereas the close-ended PREM did not

INTRODUCTION

Patient experiences are important indicators of the quality of care. According to the national health service (NHS) policies, patient experiences reflect the compassion, dignity and respect for patients during health care delivery.[1, 2] Moreover, these experiences may hold important insights for quality improvement.[3] Adequate tools to survey and analyse patient experiences are therefore essential. Patient experiences can be measured using patient-reported experience measures (PREMs), usually in the form of questionnaires.[4]

PREMs may be considered subjective, but a positive association between PREM results and other quality domains has been reported.[5] PREM scores are positively but weakly associated with patient safety and clinical effectiveness, which suggests that improving patient experiences may enhance the overall quality of care.[6, 7] Today, there are many different PREMs in use; most of them are disease or treatment specific and consist predominantly of closed-ended questions.[8-12] Some generic PREMs have been developed and are used to benchmark hospitals at a regional, national or international level.[13-17]

The increased use of PREMs is incentivised by regulatory bodies in the United Kingdom and United States of America. Frequently, PREMs are collected and analysed but translating the results into changes in clinical practice remains challenging due to organizational, professional and data-related barriers.[18-20] The lack of a quality improvement infrastructure is one of these barriers.[20] Furthermore, patient experiences are not always adopted by clinicians, because the PREM results do not provide insights relevant to their daily workflow, or because the feedback is not specific enough to allow translation into concrete action points.[3, 19] When PREM results are not translated into clear and actionable points of improvement for care providers, PREMs risk to be viewed as measurement for the sake of measurement rather than as valuable instruments for improving the underlying care.[21]

In contrast to closed-ended questions that steer a patient's feedback to a specific topic, open-ended questions enable patients to provide feedback on all aspects of care that matter to them.[22] This feature makes open-ended questions more patient-centred and yields more specific information, facilitating concrete quality improvement measures.[23] However, the analysis of free-text answers is time-consuming and too laborious to use in clinical practice.[23, 24]

Artificial intelligence (AI) techniques are able to automatically detect the topics and sentiment of patients' free text comments and help identify actionable insights out of PREMs.[25, 26]

Currently used PREMs are not ideally suited for the full exploitation of the potential of AI-techniques. First, current questionnaires often contain questions with a sentiment comprised in the question itself. (e,g: 'what went remarkably well during your stay?' or 'what could we improve?'). In addition, questions such as these invite short, monosyllabic answers, which are difficult to categorize.[25] To tackle these problems several modifications to commonly used PREMs are needed. A new AI-PREM tool has been developed and validated by Van Buchem et al.[27], with open-ended generic questions (i.e., not targeted at a specific disease, care pathway, department or healthcare centre) and suited for computer analysis by removing the sentiment from the question. The questions were focused on the Picker dimensions of patient-centred care to reduce the number of topics in an answer (e.g., What did you think about the information provision?).[27]

The primary aim of this study was to determine the added value of the AI-PREM tool compared to a conventional PREM with respect to identification of actionable points for quality improvement. The secondary aim was to assess the influence of socio-demographic determinants on AI-PREM completion and results. To do so, we have deployed the AI-PREM in a vestibular schwannoma integrated practice unit (IPU) in a vestibular schwannoma expert centre in the Netherlands.

METHODS

2.1 Context

Vestibular schwannomas are rare benign intracranial tumours, which typically cause hearing loss, tinnitus and balance disorders. A majority (52-78%) of the tumours is non-progressive. In these cases active surveillance with prolonged follow-up is usually the management strategy of choice.[28] In case of very large or progressive tumours, surgery or radiotherapy is indicated to prevent future complications such as brain stem compression. After active therapy, prolonged follow-up is warranted to detect residual or recurrent disease. Because of the long follow-up required (with or without active treatment) and near to normal life expectancy with adequate management of the tumour, patients with a vestibular schwannoma often accumulate extensive experience with healthcare professionals and centres.

2.2 Design

This descriptive case study evaluated the outcomes of an open-ended question PREM and a close-ended question PREM employed in a vestibular schwannoma IPU. A non-responder analysis was performed, the outcomes of both PREMs were analysed, and the ceiling effect was evaluated in a direct comparison. In addition, the interpretation

and the selection of actionable points of improvement by the IPU team based on these outcomes was observed. The process to come from PREM results to actionable points of improvement is reported.

The study was performed at the Leiden University Medical Centre, a tertiary university hospital, and expert centre for vestibular schwannomas in The Netherlands. At our centre, patient care is organized in an IPU, including otorhinolaryngologists, neurosurgeons, radiation oncologists and radiologists. The combination of chronic care and the multidisciplinary organization in an IPU are ideal to investigate the added value of AI-PREM for quality improvement.

2.3 Participants

This study was part of larger study on long term quality of life in vestibular schwannoma patients. For longitudinal follow-up patients who participated in 2014 in a cross-sectional survey on quality of life in vestibular schwannoma patients were re-invited for participation.[29] Using this patient group allowed the analysis of non-responders based on the data collected in 2014. In 2014, all consecutive patients who were diagnosed or treated for a unilateral vestibular schwannoma since 2003 at the IPU were eligible for inclusion. Patients under 18 years, patients with insufficient proficiency in the Dutch language to complete the questionnaires or patients with other skull base pathologies were excluded. Data collection took place between June and September 2020. The local medical research and ethics committee has waived the necessity for medical ethical approval under Dutch law and approved the study regarding data handling and privacy regulations (N19.112).

2.4 Data collection

After providing informed consent, patients were asked to complete two validated PREM questionnaires either electronically or on paper. First, participants completed the AI-PREM, consisting of five open-ended questions about their experiences with the care delivery.[27] The five questions (box 1) addressed the following themes: information provision, personal approach, collaboration, organization and other experiences, and were based on the Picker dimensions of patient-centred care.[13, 30] The free-text answers were analysed using natural language processing techniques, which divided the free-text answers into clusters of positive and negative comments. These techniques are described in more detail by Van Buchem et al.[27] The output of the AI-PREM are clusters of positive and negative comments for each of the five questions. The output was accessible in a easily intelligible dashboard. This dashboard was able to show the thematically clustered patient feedback, differentiate negative from positive clusters, and quantify the

Box 1 Questions AI-PREM [30]

- Q1: How was the provided information?
- Q2: How was the personal approach?
- Q3: How was the collaboration between healthcare professionals?
- Q4: How was the organization of care?
- Q5: What else would you like to share about your experience?

number of comments per thematic cluster. In addition, the IPU team could access the full individual patient comments the clusters were based on (as raw text).

Second, participants completed the Patient Experience Monitor (PEM) consisting of fifteen closed-ended questions about the patient's experience;[14] The PEM outcomes are proportions of patients which answered with a certain multiple choice option. For example, the proportion of the total number of respondents that trusted their physician fully.

Third, patients were asked to complete a disease-specific quality of life questionnaire of 26 items, the Penn Acoustic Neuroma Quality Of Life (PANQOL).[31, 32] Furthermore, demographic information (sex, age and education level) was acquired. Statistics Netherlands' (CBS) definition for low, middle and high education level was used, which follows the international standard classification of education.[33]

Treatment modality, tumour size at baseline, and time since diagnosis were acquired from the electronic patient records. Treatment was coded as either active surveillance, surgery or radiotherapy. Tumour size was classified according to Kanzaki et al. as intrameatal, small, moderately large, large or giant tumour.[34]

2.5 Statistical analysis

Statistical analyses were performed in R version 4.0.5 using Rstudio 1.3.959 (Rstudio, PBC, Boston).

For the demographics and non-responder analysis, means and standard deviation (sd) were calculated for normally distributed numerical variables, and medians and interquartile ranges (IQR) when not normally distributed. For categorical variables, percentages and frequencies were calculated. Demographics of non-responders, responders and one-word responders were compared using the unpaired t-test for continuous and chi-squared test for categorical variables. One word responders were defined as patients who provided a one-word answer for all open-ended questions (e.g., "well",

"fine", or "bad"). Bonferroni correction for multiple testing was used to prevent type-I errors. Incomplete questionnaires were omitted in the analysis.

The ceiling effect, a well-known feature of PREMs, was analysed using the overall experience question of the PEM. In a separate analysis, the outcome of the AI-PREM was evaluated for patients who scored >8 out of 10 on the PEM questionnaire (i.e. provided overall very positive feedback). This analysis was used to assess the capability of the AI-PREM to identify feedback that could be used for quality improvement from patients that were overall positive about their experience with the IPU.

2.6 Intervention

The results of the AI-PREM and PEM were used to identify actionable point for quality improvement. The process to analyse, interpret and translate the results are described stepwise. First, results were analysed and placed in the local context by the IPU team. This team, consisting of a deputy of each medical specialism, a researcher, a case manager and supportive staff, used their knowledge of the IPU combined with the PREM results to select feasible and effective projects.

RESULTS

In total, 536 patients provided informed consent resulting in a 62% response rate, as is shown in figure 1. Non-responders more often had a lower level of education (32% vs 44%) but a comparable mean age and male/female ratio to the responders, as shown in Table 1.

Compared to the population of vestibular schwannoma patients, the study population had a somewhat higher mean age (67.4 vs. 61.1 years) as a result of the long term follow-up. Also, the ratio of patients that received active intervention (radiotherapy or surgery) was higher (42% vs 51%), also as a result of the fact that they have been under observation for longer.

3.1 AI-PREM outcomes

The AI-PREM was completed by 507 patients, of whom 79 (16%) were one-word responders. As shown in table 1, one-word responders were more often male, two years older and had a lower education level, but these differences were not statistically significant after correcting for multiple testing. A group of 27 patients did provide informed consent but did not complete the AI-PREM and two patients were excluded because of a pathology different to vestibular schwannoma.

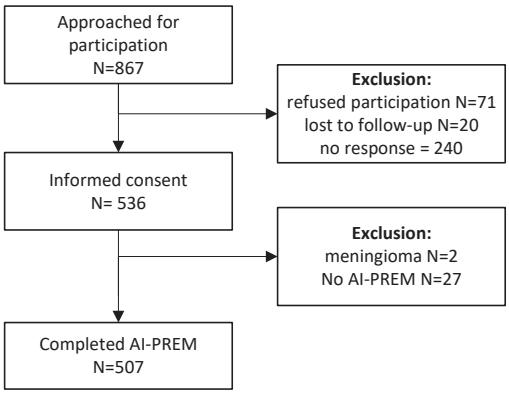


Figure 1. Flowchart study participants

Table 1. Baseline demographics

	Non-responders	Not completed	Completed	One-word answers
	N= 331	N=28	N=507	N=79
Sex (male)	49%	50%	53%	65%
Age (sd)	68.0(12.3)	69.9 (10.5)	67.4 (11.0)	69.7 (9.6)
Education level				
Low	44%	44%	32%	41%
Middle	25%	33%	30%	27%
High	31%	22%	38%	33%
Treatment				
Observation	61%*	50%	46%	49%
Surgery	26%*	29%	38%	34%
Radiotherapy	13%*	14%	13%	16%
Quality of Life (sd)	69.8 (19.8)*	66.8 (15.5)	69.2 (18.1)	70.4 (17.3)

Demographics are shown for non-responders and responders. Both incomplete and completed questionnaires are shown. One-word responders are a subcategory of completed questionnaires, in which patients completed only one-word answers, such a “good” or “bad”, on each open-ended question. Quality of life shows a disease-specific quality of life questionnaire ranging from 0-100. Higher scores indicate better quality of life. sd= standard deviation. *= data acquired in 2014

Table 2 shows the different feedback clusters of the five PREM questions including the number of comments per cluster and an example of a raw data comment. The majority of comments was classified as positive. All positive clusters contained many short or monosyllabic responses containing “well” or “fine”, which did not provide additional information or context other than the subject of the question. Negative answers were in general more detailed and contained more words. Due to the diverse nature of the negative feedback, there were more thematic clusters, each containing less individual comments. For example, three negative clusters stated that personal approach was lacking

Table 2. AI-PREM clusters open-ended question answers

Question	Information provision		Personal approach		Collaboration		Organization		Other experiences	
Clusters	*positive*	*negative*	*positive*	*negative*	*positive*	*negative*	*positive*	*negative*	*positive*	*negative*
	Well n=178	Limited* n=18	Well n=175	Insufficient n=6	Well n=215	Other hospital n=5	Well n=205	Appointment* n=34	Well* n=105	Aftercare n=14
	Clear* n=178	Lengthy n=8	Fine n=55	Reserved n=3	Fine n=98	Communication n=7	Fine n=59	Reachability n=5	Positive n=8	Waiting time* n=21
			Pleasant conversation* n=101	Personal approach n=3		Bad n=4	Well organized* n=62			
			Personal approach n=24	Limited n=3		Suboptimal n=3				
				Support n=4		Better n=3				
				Experiences* n=5		Scan n=9				
				Attention n=7		Insufficient* n=8				
Leftovers	n=3	n=0	n=5	n=0	n=10	n=1	n=17	n=0	n=8	n=0
Example raw data quotes (from cluster with *)	“The information about the disease and symptoms was clear, informative and understandable.”	“Limited. Several of my symptoms, that were in my opinion related to the tumour, were not addressed at all.”	“Excellent. Understanding and sympathetic about the symptoms. “	“The doctor’s ‘empathic capacity’ sometimes did not align with the patient’s experiences/ feelings.”		“Scheduling a follow-up scan was sometimes difficult.”	“Appointments were mostly scheduled on the same day, which was pleasant.”	“Sometimes you had to wait a long time for the appointment, no appointments scheduled on the same day. Difficult to reach by-phone.”	“The vestibular schwannoma team is well-coordinated.”	“Long waiting time for the scan results”

(N=3), limited (N=3), or insufficient (N=6). Another interesting finding was that different patients may experience certain aspects of care in a contradicting way. Therefore, the number of patients with a positive or a negative experience with the specific aspect of care was quantified, in order to put the feedback into perspective and help decide whether and which action should be taken to improve the IPU. For example, the number of patients who provided positive feedback on scheduling appointments on the same day (N=8) outnumbered those who provided negative feedback on this topic (N=2).

3.2 PEM outcomes

The PEM was completed by 490 patients. In general, the patients completed the PEM very positively and the overall experience was graded with an 8 (±1.2 sd) on a 1 to 10 point scale. For example, 95% of the patients trusted their physician, and 93% indicated they had enough time to discuss their problem with the physician. Furthermore, 93% of patients said they discussed what to do after the consultation, and 89% said they were informed about their treatment's pros and cons. The majority (87%) found the physician's explanation understandable. Only 1% indicated they could not ask questions to their consulting physician.

The question with the most negative responses concerned the waiting time in the outpatient clinic. 21% of patients indicated they had to wait >15 min. Of this group, 10% would have preferred to receive more information about the estimated waiting time.

3.3 Comparison between PREMs

Table 3 shows the AI-PREM results of patients who scored an overall experience >8 out of 10 points on the PEM questionnaire. These patients had also rather positive experiences on the AI-PREM and only a limited number of negative comments. Still, these comments provided useful and detailed information about the IPU. For example, one patient stated: *"I would have liked to hear about the treatment of vertigo with exercises sooner"*. Other patients mentioned: *"There was some misunderstanding about by whom and when I was called about an appointment."*, *"The collaboration between hospitals was poor."*, and *"I was discharged from the hospital too soon and without instructions."*

3.4 Observation of the interpretations of results

The results of the close-ended PEM questionnaire were predominantly positive, which was considered motivating information for the IPU team. However, for quality improvement these positive reactions could not be translated to action points for improvement. Conversely, the AI-PREM results provided more detailed information about the positive and negative experiences, even from patients that provided overall positive feedback. This information could be used to identify action points.

Table 3 AI-PREM results of patients with an overall PEM scores of>8/10

	Negative		Neutral		Positive	
	count	%	count	%	count	%
Information provision	3	2%	37	23%	122	75%
Personal approach	2	1%	35	22%	125	77%
Collaboration	6	4%	35	22%	121	75%
Organisation	6	4%	35	22%	121	75%
Other experiences	7	4%	90	56%	65	40%

The process to identify action points for improvement is shown in Figure 2. First, the IPU team analysed the results of the AI-PREM and explored the negative clusters of patients' experiences for potential quality improvements. The automated sentiment analysis and clustering of comments was used to identify topics of interest. These topics of interest were subsequently further explored by the IPU team through targeted evaluation of clustered patient comments (raw text). These raw texts were valued in the context of the IPU organization. When potential action points emerged they were discussed in the meeting and weighed against possible positive feedback regarding the same topic.

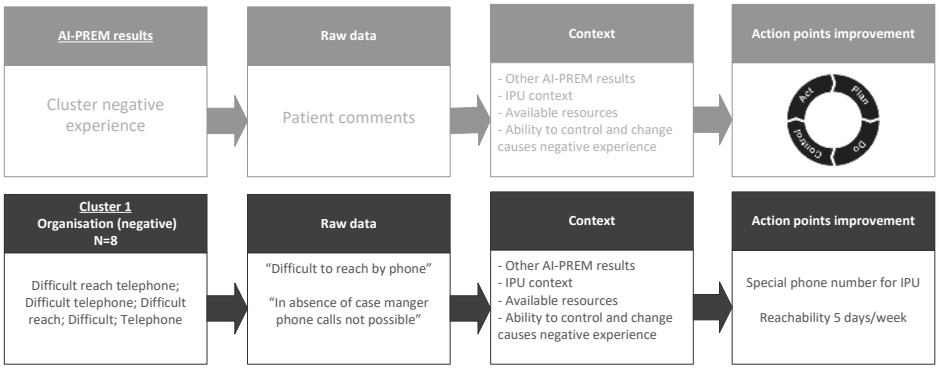


Figure 2. Process from AI-PREM results to quality improvement
The process steps from using the AI-PREM results to identify action points for quality improvement are shown in grey. The second row shows the process steps of the identified action point reachability by phone.

In all, the IPU team selected three action points for quality improvement based on actionability, feasibility and number of patients sharing the particular (negative) experience. The chosen action points were improving the reachability by phone, reducing the time between the MRI and the consultation to discuss the result and improving the communication with referring hospitals.

4. DISCUSSION AND CONCLUSION

4.1 Discussion

To our knowledge, this is the first study in which a PREM with open-ended questions is directly compared to a traditional PREM with close-ended questions. Both questionnaires allowed evaluation of patient experiences with the care provided by the vestibular schwannoma care pathway. Both questionnaires reported overall positive patients' experiences.

The PEM enabled an easy and quick quantitative analysis of the overall experience. Most results showed ceiling effects and the predefined answer categories were less suited for identification of points of improvement, especially in the context of predominantly positive experiences. The AI-PREM seemed to have a greater potential to identify actionable points for quality improvement because of the broader focus and the more detailed descriptions, especially of negative experiences. With the AI-PREM, feedback with improvement points could be obtained even from patients with very positive experiences (as judged on the PEM scores).

An essential feature determining feasibility for clinical use was the automated analysis of the open text PREMs to reduce the workload. Still, the human component in the analysis is essential to interpret the algorithm's results and combine this with the clinical context of the IPU to translate the feedback into actionable points of improvement. Furthermore, the AI-PREM combined output of quantitative and more qualitative data. This combination of sentiment scores, the number of comments per cluster and a traceback to the individual reported experience facilitated decision making for quality improvement. In contrast, the use of the structured PEM for identification of points of improvement was limited due to a small number of reported negative experiences.

The AI-PREM results showed that most comments were positive, but negative comments provided more detailed descriptions, including more context. Positive comments were more often one-word answers and generic. These findings were also described by Cunningham et al. while analysing almost 7000 open-text comments.[22] Positive comments are essential to put the negative ones into context and prioritize action points for improvement. For example, when many comments are positive about scheduling appointments, some negative comments on this cluster might be outliers, making this a less urgent target for quality improvement. In addition, positive comments can be used as motivators for the IPU team and can contribute to increasing patient safety following the Safety-II paradigm, which focuses on the things that go right rather than focusing on things that go wrong.[16, 35]

Other studies, focussing on patients narratives, have reported that the patients' comments on their experience with disease and care delivery generally provide mainly positive outcomes.[16, 17, 36] For example, the study of De Rosis et al. reported mainly positive comments which could be used for to identify positive aspects, which could be used for quality improvement by a 'learning by excellence' strategy. While this is valuable, learning by excellence in itself has a limited ability to identify actionable points for improvement. The AI-PREM presented here has the ability to show and quantify positive comments but at the same time identify points of improvement, even in the feedback of patients that are overall positive about their experience in the IPU. In doing so, a more nuanced feedback of patients on the care delivery is made possible. While we find, like previous reports, that a large majority of patients provide positive comments, we were also able to extract actionable points of improvement even from patients with generally positive feedback.

Also in research settings, generic PREMs are used to evaluate the quality improvement targeted at improving the overall patient experience.[36] Improving organizational factors for a better patient experience will not only benefit patients but has also been shown to enhance physician satisfaction.[37] However, achieving improvements in the patient experience can be challenging.[38] A large proportion of patients report high PREM scores. This ceiling effect might be caused by appreciation or social desirability bias.[39, 40] In this study, the PEM results also show this ceiling effect, which is challenging from a quality improvement perspective since these already high scores can be hard to improve on. When trying to improve patient care, focussing on overall patient satisfaction or PREM scores may therefore be less effective than evaluating the negative comments in detail. Moreover, this study shows that even patients with a positive overall experience (as reported in the PEM) may still have feedback indicating points of improvement (identified with the AI-PREM). The AI-PREM design allows for an in-depth analysis of the comments by grouping them together in clusters based on sentiment and similar word content. Consequently, the actual remarks concerning a certain topic made by individual patients can be accessed, providing all necessary detail, without manually going through all questionnaires to extract information about the topic at hand. This approach, which yields both quantitative and qualitative data from free-text answers, saves time yet allows patients to comment freely on their experience with all aspects of care, detailed analysis of their feedback and identification of specific points of improvement.

A potential problem of using PREMs for quality improvements is a selection bias of the patients who complete the PREMs. When the responders are not a random sample of the total patient population the risk for inadequately aimed quality optimisations exists. Younger patients and black, indigenous and people of colour tend to report less positive

patient experiences.[41, 42] So it is important to include answers of these groups in the analysis for quality improvement. The non-responder analysis showed a larger proportion of lower education level in this group. There were no age differences, but one-word responders were on average slightly elder. These aspects should be considered when interpreting the PREM results to prevent nonresponse errors.[43]

In addition, open-ended PREMs might reflect the a priori expectations and perceptions of care. When the provided care meets the expectations, patients might not provide feedback but they probably will when the experience is worse or much better than their expectations. This phenomenon is especially important since different populations have different expectations of care delivery.[44, 45] The evolution from patient satisfaction (e.g., how would you rate the information you received about your treatment?) towards experience (e.g., did you receive information about your treatment?) has mitigated the risk of such bias.[45] However, open-ended questions in structured PREMs are often focussed on patient satisfaction (e.g., "What went remarkably well during your stay?"). The AI-PREM questions focus more on the experience and reduce but not neutralize the risk of expectation bias.

In this study, a patient population was selected that had already participated in previous research. These dedicated participants might introduce some selection bias. When collecting the PREMs prospectively, the response rate might, therefore, be lower. Another limitation was the prolonged recall period since the last visit to the hospital in this research. The period exceeded the 4-6 weeks used in the PEM validation study.[14] This prolonged period might have limited the output of the PREMs.[2] However, the comparison between the two PREMs was not affected since both questionnaires were completed simultaneously.

4.2 Experiences of deployment in a vestibular schwannoma IPU

The IPU team used the PREM results to identify actionable points for quality improvement. This entailed a process of interpretation of the PREM results and analysing them in order to use them to improve clincal practice. Important parameters during the IPU team discussions were the quantitative results and the positive feedback clusters. The quantitative information (how many patients shared the same view) was useful in determining the extent of the problem. However, the positive feedback was essential too, for putting certain negative comments into perspective and prioritizing and focusing actions on improving the care delivery. Taking action based on the negative comments only could mistakenly alter aspects of care that provided a positive experience for most patients.

In addition, the potential of the IPU to improve or change the underlying causes of the negative experience was discussed. For example, a negative patient experience about a lack of parking space is beyond the control of the IPU, but the communication about the appointments is within the sphere of influence of the IPU. When potential action points were within the sphere of influence, the available resources needed to perform an improvement cycle were identified to see whether an improvement cycle was feasible. Finally, the IPU team decided to start a plan, do, check, act cycle.

4.3 Innovation

With the growing interest in patient-centeredness of care comes a growing need to adequately assess the patient experience with care delivery. The AI-PREM may be a tool that allows patients to freely comment on their experience yet is economic with the time and effort invested by healthcare professionals to analyse the feedback, although the time and effort invested by patients to complete the AI-PREM should also be considered. To make the efforts of patients worthwhile, PREMs should be used to improve care delivery, rather than as an administrative requirement. Future research should evaluate the applicability of the AI-PREM in different clinical settings. Because of the generic nature of the AI-PREM questionnaire, it seems likely to be of value in a multitude of different diseases, care pathways, or healthcare centres. In addition, the ability of the AI-PREM to detect longitudinal changes in the quality of care and/or the effect of measures to improve the quality of care may be the subject of future research.

4.4 Conclusion

Patient experiences are an essential aspect of quality of care. This study showed the added value of open-ended PREM questions in assessing patient experiences. The AI-PREM provided insights into both positive and negative experiences and allowed the detection of actionable targets for quality improvement in an IPU. Because of its automated analysis and readily accessible results, the evaluation of the patient experience with the vestibular schwannoma care pathway could be performed by IPU clinicians and translated into action points relevant to context of the clinical IPU.

REFERENCES

[1] Department of Health, High quality care for all: NHS next stage review final report., in: D.o. Health (Ed.) The Stationery Office, London, 2008.

[2] M.P. Manary, W. Boulding, R. Staelin, S.W. Glickman, The Patient Experience and Health Outcomes, N. Engl. J. Med. 368(3) (2013) 201-203.

[3] H. Gleeson, A. Calderon, V. Swami, J. Deighton, M. Wolpert, J. Edbrooke-Childs, Systematic review of approaches to using patient experience data for quality improvement in healthcare settings, BMJ Open 6(8) (2016) e011907.

[4] C. Bull, J. Byrnes, R. Hettiarachchi, M. Downes, A systematic review of the validity and reliability of patient-reported experience measures, Health Serv. Res. 54(5) (2019) 1023-1035.

[5] F. Greaves, A.K. Jha, Quality and the curate's egg, BMJ Quality & Safety 23(7) (2014) 525-527.

[6] N. Black, M. Varaganum, A. Hutchings, Relationship between patient reported experience (PREMs) and patient reported outcomes (PROMs) in elective surgery, BMJ Quality & Safety 23(7) (2014) 534-542.

[7] C. Doyle, L. Lennox, D. Bell, A systematic review of evidence on the links between patient experience and clinical safety and effectiveness, BMJ Open 3(1) (2013) e001570.

[8] M.B. Rivara, T. Edwards, D. Patrick, L. Anderson, J. Himmelfarb, R. Mehrotra, Development and Content Validity of a Patient-Reported Experience Measure for Home Dialysis, Clin. J. Am. Soc. Nephrol. 16(4) (2021) 588-598.

[9] L. Zinckernagel, N. Schneekloth, A.-D.O. Zwisler, A.K. Ersbøll, M.H. Rod, P.D. Jensen, H. Timm, T. Holmberg, How to measure experiences of healthcare quality in Denmark among patients with heart disease? The development and psychometric evaluation of a patient-reported instrument, BMJ Open 7(10) (2017) e016234.

[10] R.M. Taylor, L.A. Fern, A. Solanki, L. Hooker, A. Carluccio, J. Pye, D. Jeans, T. Frere–Smith, F. Gibson, J. Barber, R. Raine, D. Stark, R. Feltbower, S. Pearce, J.S. Whelan, Development and validation of the BRIGHTLIGHT Survey, a patient-reported experience measure for young people with cancer, Health and Quality of Life Outcomes 13(1) (2015).

[11] A. Bosworth, M. Cox, A. O'Brien, P. Jones, I. Sargeant, A. Elliott, M. Bukhari, Development and Validation of a Patient Reported Experience Measure (PREM) for Patients with Rheumatoid Arthritis (RA) and other Rheumatic Conditions, Curr. Rheumatol. Rev. 11(1) (2015) 1-7.

[12] N. Bobrovitz, M. Santana, T. Kline, J. Kortbeek, H.T. Stelfox, Prospective cohort study protocol to evaluate the validity and reliability of the Quality of Trauma Care Patient-Reported Experience Measure (QTAC-PREM), BMC Health Serv. Res. 13(1) (2013) 98.

[13] C. Jenkinson, The Picker Patient Experience Questionnaire: development and validation using data from in-patient surveys in five countries, Int. J. Qual. Health Care 14(5) (2002) 353-358.

[14] C.M. Bastemeijer, H. Boosman, L. Zandbelt, R. Timman, D. De Boer, J.A. Hazelzet, Patient Experience Monitor (PEM): The Development of New Short-Form Picker Experience Questionnaires for Hospital Patients with a Wide Range of Literacy Levels</p>, Patient Related Outcome Measures Volume 11 (2020) 221-230.

[15] L.A. Giordano, M.N. Elliott, E. Goldstein, W.G. Lehrman, P.A. Spencer, Development, implementation, and public reporting of the HCAHPS survey, Med. Care Res. Rev. 67(1) (2010) 27-37.

[16] S. De Rosis, D. Cerasuolo, S. Nuti, Using patient-reported measures to drive change in healthcare: the experience of the digital, continuous and systematic PREMs observatory in Italy, BMC Health Serv. Res. 20(1) (2020).

[17] I. Corazza, K.J. Gilmore, F. Menegazzo, V. Abols, Benchmarking experience to improve paediatric healthcare: listening to the voices of families from two European Children's University Hospitals, BMC Health Serv. Res. 21(1) (2021).

[18] A. Decourcy, E. West, D. Barron, The National Adult Inpatient Survey conducted in the English National Health Service from 2002 to 2009: how have the data been used and what do we know as a result?, BMC Health Serv. Res. 12(1) (2012) 71.

[19] A. Coulter, L. Locock, S. Ziebland, J. Calabrese, Collecting data on patient experience is not enough: they must be used to improve care, BMJ 348(mar26 1) (2014) g2225-g2225.

[20] E. Davies, Hearing the patient's voice? Factors affecting the use of patient survey data in quality improvement, Quality and Safety in Health Care 14(6) (2005) 428-432.

[21] M. Kunneman, V.M. Montori, N.D. Shah, Measurement with a wink, BMJ Quality & Safety 26(10) (2017) 849-851.

[22] M. Cunningham, M. Wells, Qualitative analysis of 6961 free-text comments from the first National Cancer Patient Experience Survey in Scotland, BMJ Open 7(6) (2017) e015726.

[23] E. Riiskjaer, J. Ammentorp, P.E. Kofoed, The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective, Int. J. Qual. Health Care 24(5) (2012) 509-516.

[24] J. Garcia, J. Evans, M. Reshaw, ``Is There Anything Else You Would Like to Tell Us'' – Methodological Issues in the Use of Free-Text Comments from Postal Surveys, Quality & Quantity 38(2) (2004) 113-125.

[25] S.A. Cammel, M.S. De Vos, D. Van Soest, K.M. Hettne, F. Boer, E.W. Steyerberg, H. Boosman, How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach, BMC Med. Inform. Decis. Mak. 20(1) (2020).

[26] C. Arditi, D. Walther, I. Gilles, S. Lesage, A.-C. Griesser, C. Bienvenu, M. Eicher, I. Peytremann-Bridevaux, Computer-assisted textual analysis of free-text comments in the Swiss Cancer Patient Experiences (SCAPE) survey, BMC Health Serv. Res. 20(1) (2020).

[27] M.M. Van Buchem, O.M. Neve, I.M.J. Kant, E.W. Steyerberg, H. Boosman, E.F. Hensen, Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (AI-PREM), BMC Med. Inform. Decis. Mak. 22(1) (2022).

[28] M.L. Carlson, M.J. Link, Vestibular Schwannomas, N. Engl. J. Med. 384(14) (2021) 1335-1348.

[29] G. Soulier, B.M. Van Leeuwen, H. Putter, J.C. Jansen, M.J.A. Malessy, P.P.G. Van Benthem, A.G.L. Van Der Mey, A.M. Stiggelbout, Quality of Life in 807 Patients with Vestibular Schwannoma: Comparing Treatment Modalities, Otolaryngology–Head and Neck Surgery 157(1) (2017) 92-98.

[30] M. Gerteis, S. Edgman-Levitan, J. Daley, T.L. Delbanco, Through the Patient's Eyes: Understanding and Promoting Patient-Centered Care, Jossey-Bass, San Francisco, 1993.

[31] B.M. van Leeuwen, J.M. Herruer, H. Putter, J.C. Jansen, A.G. van der Mey, A.A. Kaptein, Validating the Penn Acoustic Neuroma Quality Of Life Scale in a sample of Dutch patients recently diagnosed with vestibular schwannoma, Otol. Neurotol. 34(5) (2013) 952-7.

[32] B.T. Shaffer, M.S. Cohen, D.C. Bigelow, M.J. Ruckenstein, Validation of a disease-specific quality-of-life instrument for acoustic neuroma, The Laryngoscope 120(8) (2010) 1646-1654.

[33] Statistics Netherlands, Standaard Onderwijsindeling 2016, Den Haag, 2017.

[34] J. Kanzaki, M. Tos, M. Sanna, D.A. Moffat, E.M. Monsell, K.I. Berliner, New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma, Otol. Neurotol. 24(4) (2003) 642-8; discussion 648-9.

[35] J. Braithwaite, R.L. Wears, E. Hollnagel, Resilient health care: turning patient safety on its head, Int. J. Qual. Health Care 27(5) (2015) 418-420.

[36] C.M. Bastemeijer, H. Boosman, H. Van Ewijk, L.M. De Jong-Verweij, L. Voogt, J. Hazelzet, Patient experiences: a systematic review of quality improvement interventions in a hospital setting, Patient Related Outcome Measures Volume 10 (2019) 157-169.

[37] N. Golda, S. Beeson, N. Kohli, B. Merrill, Analysis of the patient experience measure, J. Am. Acad. Dermatol. 78(4) (2018) 645-651.

[38] E. Wong, F. Mavondo, L. Horvat, L. McKinlay, J. Fisher, Victorian healthcare experience survey 2016–2018; evaluation of interventions to improve the patient experience, BMC Health Serv. Res. 21(1) (2021).

[39] I.I. Kleiss, J.T. Kortlever, P. Karyampudi, D. Ring, L.E. Brown, L.M. Reichel, M.D. Driscoll, G.A. Vagner, A comparison of 4 single-question measures of patient satisfaction, J. Clin. Outcomes Manag. 27 (2020).

[40] A.A. Salman, B.J. Kopp, J.E. Thomas, D. Ring, A. Fatehi, What Are the Priming and Ceiling Effects of One Experience Measure on Another?, Journal of Patient Experience 7(6) (2020) 1755-1759.

[41] J.L. Campbell, Age, gender, socioeconomic, and ethnic differences in patients' assessments of primary health care, Qual. Health Care 10(2) (2001) 90-95.

[42] G. Lyratzopoulos, M. Elliott, J.M. Barbiere, A. Henderson, L. Staetsky, C. Paddison, J. Campbell, M. Roland, Understanding ethnic and other socio-demographic differences in patient experience of primary care: evidence from the English General Practice Patient Survey, BMJ Quality & Safety 21(1) (2012) 21-29.

[43] T.P. Johnson, Response Rates and Nonresponse Errors in Surveys, JAMA 307(17) (2012) 1805.

[44] K.G. Poole, Patient-Experience Data and Bias — What Ratings Don't Tell Us, N. Engl. J. Med. 380(9) (2019) 801-803.

[45] F. Ahmed, J. Burt, M. Roland, Measuring Patient Experience: Concepts and Methods, The Patient - Patient-Centered Outcomes Research 7(3) (2014) 235-241.

Fully Automated 3D Vestibular Schwannoma Segmentation with and without Gadolinium Contrast: A Multicenter, Multivendor Study

Olaf Neve*
Yunjie Chen*
Qian Tao
Stephan Romeijn
Nick de Boer
Mark Kruit
Bouwdewijn Lelieveldt
Jeroen Jansen
Erik Hensen
Berit Verbist
Marius Staring

* contributed equally to this work.

Radiol Artif Intell. 2022 Jun 22;4(4):e210300
doi: 10.1148/ryai.210300

ABSTRACT

Purpose

To develop automated vestibular schwannoma measurements on contrast-enhanced T1- and T2-weighted MRI.

Material and methods

MRI data from 214 patients in 37 different centers was retrospectively analyzed between 2020-2021. Patients with hearing loss (134 vestibular schwannoma positive [mean age ± SD, 54 ± 12 years; 64 men], 80 negative) were randomized to a training and validation set and an independent test set. A convolutional neural network (CNN) was trained using five-fold cross-validation for two models (T1 and T2). Quantitative analysis including Dice index, Hausdorff distance, surface-to-surface distance (S2S), and relative volume error were used to compare the computer and the human delineations. Furthermore, an observer study was performed in which two experienced physicians evaluated both delineations.

Results

The T1-weighted model showed state-of-the-art performance with a mean S2S distance of less than 0.6 mm for the whole tumor and the intrameatal and extrameatal tumor parts. The whole tumor Dice index and Hausdorff distance were 0.92 and 2.1 mm in the independent test set. T2-weighted images had a mean S2S distance less than 0.6 mm for the whole tumor and the intrameatal and extrameatal tumor parts. Whole tumor Dice index and Hausdorff distance were 0.87 and 1.5 mm in the independent test set. The observer study indicated that the tool was comparable to human delineations in 85-92% of cases.

Conclusion

The CNN model detected and delineated vestibular schwannomas accurately on contrast-enhanced T1 and T2-weighted MRI and distinguished the clinically relevant difference between intrameatal and extrameatal tumor parts.

1. INTRODUCTION

Vestibular schwannomas are rare, benign intracranial tumors arising from the neurilemma of the vestibular nerve. Initial symptoms usually comprise hearing loss, tinnitus, and balance disturbance. Approximately 60% of tumors show no or minimal progression over time, while 40% are either very large at presentation or show progression during follow-up.[1] Small to medium-sized tumors are not life-threatening and are generally conservatively managed, at least initially, using surveillance with repeated MRIs. Conversely, patients with large tumors at presentation or with tumors that progress during follow-up may need intervention through either radiotherapy or surgery. Currently, there are no reliable predictors for tumor progression.

Currently, tumor progression is determined based on the extrameatal manual diameter measurements on subsequent MRIs.[2] However, these two-dimensional (2D) measurements have considerable error, resulting in inter- and intraannotator differences of 10-40%.[3-5] The more accurate three-dimensional (3D) volume measurements have not been widely applied in clinical practice since these measurements are time-consuming.[3-6]

To address this problem, several automated segmentation tools have been developed in recent years.[7, 8, 9] The reported tools were trained for volume measurement of vestibular schwannoma on gadolinium-enhanced T1-weighed MRIs and sometimes additional T2-weighted MRIs. These tools are increasingly based on deep learning methods, which yield state-of-the-art performance in many vision tasks including medical image segmentation. Deep convolutional neural networks (CNNs), particular the UNet architecture, can reach expert-level performance in various organ segmentation tasks from clinical MRI.[8] Although many variants of the UNet have been proposed and demonstrated task-specific improvements, recent insights suggest that rather than the architecture, careful selection of the hyperparameters and training strategy can have an important effect on performance.[9] The no-new-UNet framework, abbreviated nnUNet, indeed demonstrated this for several organs and imaging modalities.[10, 11] As such, we propose application of nnUNet to address vestibular schwannoma segmentation in our clinical setting.

This study aimed to develop a deep learning CNN model to automatically detect and segment vestibular schwannoma in 3D from T2-weighted and gadolinium-enhanced T1-weighted MRI, acquired from multiple centers using different MRI scanners and scan protocols. We additionally carried out a carefully designed observer study, based on the concept that the radiologists' visual observation of the segmentation results can be a direct, important evaluation of segmentation quality. In addition to conventional metrics, the observer study highlights the applicability of our model in a clinical setting.

1. MATERIALS AND METHODS

This retrospective study was performed at the Leiden university Medical Center, a tertiary referral center for vestibular schwannoma in 2020-2021. The institutional review board approved the study protocol (G19.115) and waived the obligation to obtain informed consent.

2.1 Patients and Data

In total, 214 patients who underwent an MRI examination because of hearing loss were included in the study, with 134 patients who were vestibular schwannoma-positive (mean age, 54 ± [SD] 12 years; 64 men) and 80 who were vestibular schwannoma-negative. Vestibular schwannoma patient selection included a wide spectrum of patient and tumor characteristics such as patient age, sex, tumor size and tumor consistency. All positive patients were adults with a unilateral vestibular schwannoma, and at least one gadolinium-enhanced T1-weighted MRI. High-resolution T2-weighted images were available in 112 patients. MRIs post-surgery or irradiation were excluded. Available MRI examinations were originally acquired in 37 different hospitals on 12 different MRI scanners from 3 major MRI vendors. The MRIs of negative cases, included to optimize detection performance, were solely acquired at the LUMC of adult patients with hearing loss prior to cochlear implantation, and had no demographic data available due to prior anonymization. Patients' characteristics and technical information is shown in Table 1. In positive cases, the intra and extrameatal components[2] and the whole tumor were manually delineated by two annotators independently (ON M.D. 3 years of experience and SR technical physician, 2 years of experience) on the gadolinium enhanced T1-weighted MRI, supervised and when necessary corrected by a senior head-and-neck radiologist (BV). Two senior radiologists with 18 (MK) and 21 (BV) years of experience trained both annotators. Delineation was performed using Vitrea software v7.14.2.227 (Vital Images Inc., Minnetonka, MN, USA). The delineation was automatically propagated to T2-weighted MRI after rigid image registration using elastix.[12, 13] The complete data set was split into a training and validation set (80% from 26 centers), and an independent test set (20% from 11 different centers) on which the model was not trained, see Figure 1 for details. This was done to mimic clinical deployment where new cases may be slightly different from the data seen in the training phase and possibly bear an unknown distribution shift.[14]

Furthermore, the publicly available data set by Shapey et al. was used as additional external test of the contrast-enhanced T1-weighted model (n=242).[15] This dataset contained 47 post-surgery scans, which were omitted from the analysis.

Table 1. Patient and Technical Characteristics

Patients with vestibular schwannoma	**Value**	
N	134	
Age (y), mean (SD)	54 (12)	
Sex, men	64 (48%)	
Cystic component	63 (47%)	
Tumor size		
Intrameatal	28 (21%)	
Small (0-10mm)	19 (14%)	
Medium (11-20mm)	26 (19%)	
Moderately large (21-30mm)	24 (18%)	
Large (31-40mm)	24 (18%)	
Giant (>40mm)	13 (10%)	
Technical MRI features	**Contrast-enhanced T1**	**T2**
	Median (range)	**Median (range)**
N	134	112
In-plane resolution (mm)	0.35x0.35 (0.27x0.27 - 1.0x1.0)	0.29x0.29 (0.23x0.23 - 0.70x0.70)
In-plane matrix	400x400 (256x208 - 560x560)	512x512 (256x192 - 768x652)
TE (ms)	9 (2.38 - 20)	200 (1.53 - 297)
TR (ms)	602.10 (8.76 - 2200)	2400 (4.47 - 5000)
Slice thickness (mm)	1.0 (0.9 - 5.0)	0.6 (0.5 – 1.8)

Note.—Data presented as number of patients (percentage), unless otherwise noted. TE = echo time, TR = repetition time, SD = standard deviation

2.2 CNN Architecture and Training

NnUNet is a deep learning-based segmentation method that automatically selects one of three network architectures, includes pre-processing and post-processing methods, and performs automatic tuning of hyperparameters.[10] In this study, a 3D U-net with five encoder and decoder layers was selected, using randomly cropped 3D image patches of size 320x320x20 voxels as network input during training. The network was trained as a multi-class segmentation task to automatically segment both the intra and extrameatal component of the tumor. Two 3D nnUNets were trained, one for contrast-enhanced T1, and one for T2-weigthed MRI, from scratch with He initialization. Five-fold cross-validation was used, generating five models that were merged by averaging the softmax scores. To deal with multi-center settings, z-scoring normalization was performed to each image independently. All the training images were then resampled to the median spacing of the training dataset using third-order spline interpolation. Training was performed on an NVIDIA Tesla V100 graphics processing unit with 16GB memory using the PyTorch (v1.7.1) library.

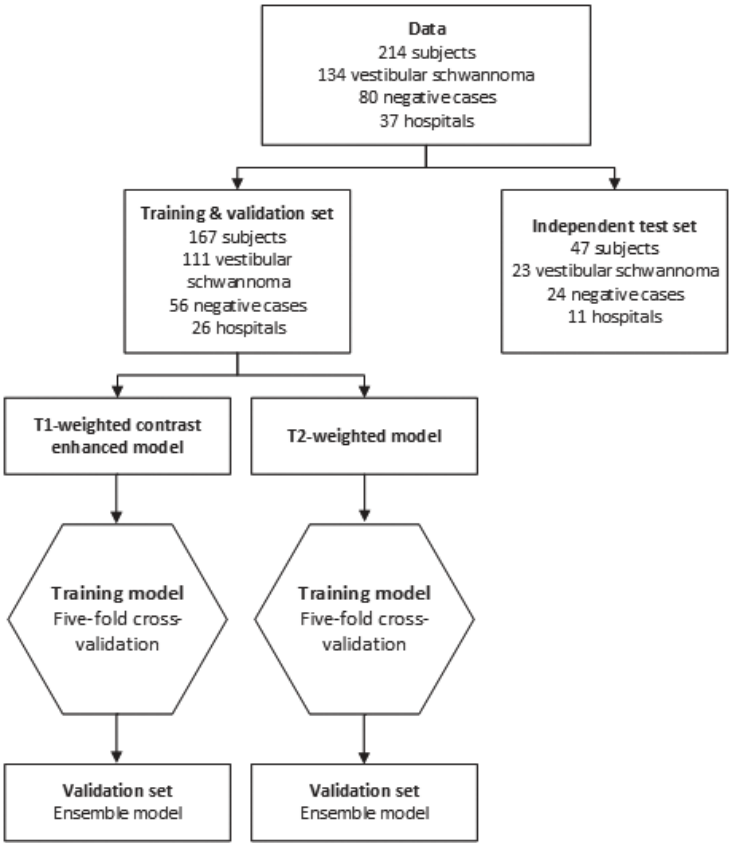


Figure 1. Flowchart of data. Patients were randomized to the training and validation set (80%) and the independent test set (20%). Positive cases were randomized based on the hospital where the scan was acquired, so the independent test set contained data of 11 hospitals that were not used to train the algorithm. For training and validation, five-fold cross-validation was used. The average of the five models is the ensemble model. This ensemble model was evaluated in the independent test set.

2.3 Observer Study

An observer study was performed to test whether the CNN could perform as well as human delineation on contrast-enhanced T1-weighted images. The T1-weighted annotations were propagated to T2-weighted MRI; therefore, the observer study was only conducted for the T1-weighted images. A user interface was created (Fig. 2), showing a gadolinium-enhanced T1-weighted image and the registered T2-weighted image in the top row and the human and automatic delineation in random order on the bottom row, projected on the gadolinium-enhanced T1-weighted MRI. Observers were able to scroll through the MRI, manually adjust its brightness and contrast, and toggle the segmentations on and off for optimal assessment. The observers were a head-and-neck radiologist (BV) and a skull base otorhinolaryngologist (EH, 18 years of experience), blinded for case information and delineation type (human or automated). The observers were

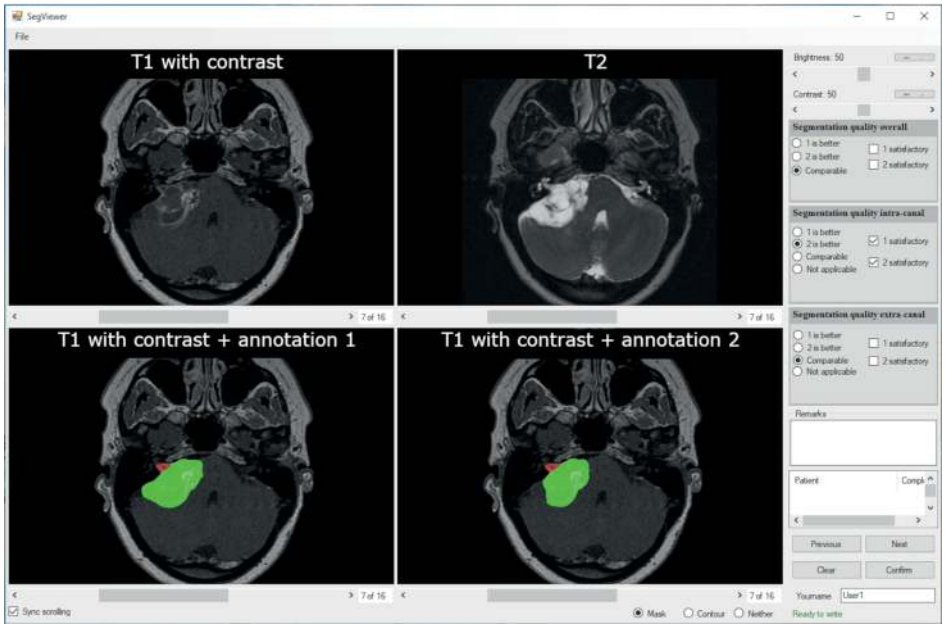


Figure 2. Observer study interface. The top row shows the clean, gadolinium-enhanced T1-weighted MRI and T2-weighted MRI. The bottom row shows the convolutional neural network and human annotations, randomized to left and the right pane, respectively. The multiple-choice questions for each observer are shown at the right side of the interface. The observers could additionally add free text comments.

asked to rate and compare the two delineations by answering two separate questions about the intra- and extrameatal part and the whole tumor: (1) Which delineation is better (annotation 1, annotation 2, or comparable), (2) Is the annotation quality satisfactory (yes or no). In a consensus meeting, cases in which observers did not agree were discussed. The consensus results are presented in section 3.5.

2.4 Testing and Statistical Analysis

All test images were resampled in the same way as the training data, and a sliding window approach was used to predict images with a window size of 320x320x10 voxels, which is the same as the network's input size. The step size is half of the window size, and a Gaussian weighted function was applied in aggregating the predictions. To eliminate false detection, connected component-based post-processing was performed. Only the largest connected component in the predictions was kept. Tumor detection by the CNN was defined as at least one voxel being detected. The performance was evaluated using the Dice index measuring overlap of the delineations, 95th percentile Hausdorff distance indicating the maximum distance between delineations, surface-to-surface (S2S) distance indicating the mean distance between delineations, and the relative volume error (RVE) indicating the difference in volume in percentage. One of the annotator's

(ON annotator 1) delineations were used for training and quantitative evaluation. The results were plotted in box-and-whisker plots. Furthermore, inter-annotator variability was investigated. Differences between the prediction performance of each annotator and the inter-annotator variabilities were tested using Wilcoxon signed-rank test. In addition, a post hoc analysis was conducted of T1-model performance with respect to tumor size, according to the classification by Kanzaki et al.2 To avoid group sizes that were too small per category, the validation and test set were pooled and a Kruskal Wallis test was performed. P-values < .05 were considered statistically significant. Observer agreement before the consensus meeting on satisfactory degree for segmentation and human delineation was expressed as percentage agreement. All analyses were performed in Python (v3.8.2) with NumPy (v1.20.2), SciPy (v1.3.3) and the sklearn (v0.23.2) library.

3. RESULTS

The CNN detected tumors with 100% sensitivity and 99.1% specificity for the validation set and 100% sensitivity and 100% specificity for the test set. The algorithm was able to calculate the segmentation with a median runtime of 78 seconds per patient.

3.1 Performance with Contrast-enhanced T1-weighted MRI

The results of the CNN on contrast-enhanced T1-weighted MRI are shown in Table 2 and Figure 3A. S2S distance of the whole tumor is 0.31mm ± [SD] 0.36 and 0.47 mm ± 0.67 in the validation set and independent test set, respectively. These S2S distances are around the in-plane voxel size and lower than the slice thickness. The whole tumor Hausdorff distance in the independent test set was 2.10mm ± 3.34, and 1.34mm ± 0.84 and 2.18mm ± 3.43, in the intra- and extrameatal parts, respectively. All the median Hausdorff distances were below the 2 mm threshold, which is often used in clinical practice to define 2D growth.[1] T1 model performance on the independent test set was comparable to the results in the validation set, indicating robust external validity. Remarkably, the independent test set had higher mean Hausdorff properties compared to the median due to two outliers (cystic tumor) in the test set which influenced the Hausdorff distance and its standard deviation. Dice indices for the whole tumor were above 0.91±0.10 and 0.92±0.05 in both sets, and RVE 7.6±4.9 and 10.2±9.1 , with lower values for the intra- and extrameatal parts of the tumor due to the sensitivity of Dice and RVE to small volumes. Figure 4 shows some examples of the T1 model compared with annotator 1.

Table 2. Quantitative Results of the Contrast-enhanced T1-weighted Model

(a) Validation set								
	Dice		**95% Hausdorff (mm)**		**S2S (mm)**		**RVE (%)**	
	mean ± SD	**median**	**mean ± SD**	**median**	**mean ± SD**	**Median**	**mean ± SD**	**median**
Whole tumor	0.91 ± 0.10	0.93	1.13 ± 1.45	1.00	0.31 ± 0.36	0.24	7.59 ± 8.10	4.88
Intrameatal	0.78 ± 0.21	0.85	1.26 ± 0.78	1.00	0.31 ± 0.20	0.26	19.7± 43.5	11.5
Extrameatal	0.83 ± 0.26	0.93	1.43 ± 1.67	1.00	0.41 ± 0.43	0.31	12.0 ± 21.6	4.94
(b) Independent test set								
	Dice		**95% Hausdorff (mm)**		**S2S (mm)**		**RVE (%)**	
	mean ±SD	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**
Whole tumor	0.92 ± 0.05	0.93	2.10 ± 3.34	1.00	0.47 ± 0.67	0.36	10.2 ± 9.1	7.1
Intrameatal	0.81 ± 0.08	0.81	1.34 ± 0.84	1.12	0.37 ± 0.23	0.32	14.7 ± 14.8	6.8
Extrameatal	0.89 ± 0.12	0.93	2.18 ± 3.43	1.00	0.52 ± 0.68	0.37	12.1 ± 16.9	6.5
(c) Publicly available dataset by Shapey et al.								
	Dice		**95% Hausdorff (mm)**		**S2S (mm)**		**RVE (%)**	
	mean ±SD	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**
Whole tumor	0.88 ± 0.04	0.88	1.31 ± 0.22	1.30	0.39 ± 0.12	0.37	27.6 ± 11.9	26.1

Dice index, Hausdorff distance, surface-to-surface distance (S2S) and relative volume error (RVE) of the model compared with annotator 1 in the (a) validation set, (b) independent test set, and (c) publicly available data set by Shapey et al. The publicly available data set seems to have structurally smaller ground truths, as can be seen in Fig. D in the supplemental material. SD = standard deviation

The CNN model, when applied to the publicly available dataset of Shapey et al., performed at the same level as with the independent test set, with a mean Dice index of 0.88±0.04, a mean Hausdorff distance of 1.31mm±0.22, a mean S2S distance of 0.39 mm±0.12, and an RVE of 26%±11.9.

3.2 Performance with T2-weighted MRI

The results of the whole tumor and the intra- and extrameatal parts are summarized in Table 3 and Figure 3B. S2S distances ranged between 0.46±0.28 and 1.00 mm ±3.75for all tumor parts in both data sets. Hausdorff distance of the whole tumor in the validation set was 3.12 mm ±9.28, with a smaller value in the independent test set (1.52 mm ±0.76). Whole tumor Dice indices were 0.82±0.19 and 0.87± 0.06 and RVE values ranged from 12.1% ± 10.8 and 24.5% ±98.8 in both data sets. Intrameatal tumors had worse Dice indices and RVE0.69±0.23 and 0.74±0.08 and 14.5%± 18.7 and % ± 21.2, respectively, likely due to the low contrast between the tumor and adjacent petrous bone in T2-weighted images. Overall T2 performance was slightly degraded compared to post-contrast T1. However, S2S distances below 1 mm indicate acceptable performance.

Table 3. Quantitative Results of the T2-weighted Model

	Dice		95% Hausdorff (mm)		S2S (mm)		RVE (%)	
(a) Validation set								
	mean ± SD	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**
Whole tumor	0.82 ± 0.19	0.87	3.12 ± 9.28	1.27	1.00 ± 3.75	0.42	24.5 ± 98.9	7.60
Intrameatal	0.69 ± 0.23	0.78	1.60 ± 0.95	1.20	0.46 ± 0.28	0.40	14.5 ± 18.7	8.39
Extrameatal	0.77 ± 0.28	0.88	2.70 ± 3.19	1.67	0.82 ± 1.01	0.54	30.9 ± 73.3	18.5
(b) Independent test set								
	Dice		**95% Hausdorff (mm)**		**S2S (mm)**		**RVE (%)**	
	mean ± SD	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**
Whole tumor	0.87 ± 0.06	0.89	1.52 ± 0.76	1.21	0.54 ± 0.31	0.47	12.1 ± 10.8	9.01
Intra meatal	0.74 ± 0.08	0.74	1.64 ± 0.59	1.50	0.52 ± 0.20	0.50	12.6 ± 21.2	5.27
Extrameatal	0.85 ± 0.17	0.89	1.60 ± 0.92	1.14	0.56 ± 0.33	0.42	22.3 ± 14.9	20.0

Dice index, Hausdorff distance, surface-to-surface distance (S2S) and relative volume error (RVE) of the model compared with the annotator 1 in the (a) validation set and (b) independent test set. SD = standard deviation

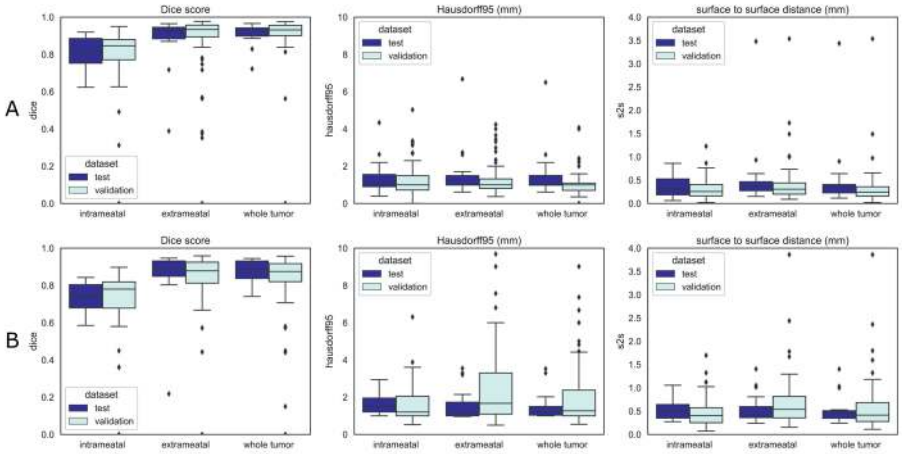


Figure 3. Quantitative boxplots of convolutional neural network tumor segmentation performance. The Dice 95% Hausdorff (Hausdorff95) distance, and surface-to-surface distance (S2S) measures are shown from left to right. (A) Boxplots of the contrast-enhanced T1 model. (B)Results of the T2-weighted model. Validation set results are shown in sky blue and independent test set results in dark blue.

3.3 Inter-annotator Variability

Comparisons between the T1-weighted model and the two annotators and between the two annotators are shown in Table 4 and Figure 5. The comparison between both annotators shows the whole tumor inter-annotator variability, resulting in a Dice index around 0.91 and RVE of 7-9%. When the model was compared to each annotator in both datasets, S2S distances were similar and below 0.5 mm. The model was trained on annotator 1, but the results compared with annotator 2 are similar for all quantitative measures.

Table 4. Comparison of the Model with Annotators and Inter-annotator Variability

	(a) Validation set											
	Dice			95% Hausdorff (mm)			S2S (mm)			RVE(%)		
	mean ± SD	p-value	median	mean ± SD	p-value	median	mean ± SD	p-value	median	mean ± SD	p-value	median
CNN – ann 1	0.91 ± 0.10	<.001	0.93	1.13 ± 1.45	<.001	1.00	0.31 ± 0.36	<.001	0.24	7.59 ± 8.10	.21	4.88
CNN – ann 2	0.90 ± 0.11	.40	0.92	1.33 ± 1.52	.18	1.00	0.36 ± 0.36	.58	0.31	10.1 ± 9.8	.35	7.1
ann 1 – ann 2	0.91 ± 0.05		0.92	1.27 ± 0.82		1.00	0.34 ± 0.20		0.31	9.01 ± 9.14		6.40
	(b) Independent test set											
	Dice			95% Hausdorff (mm)			S2S (mm)			RVE(%)		
	mean ± SD	p-value	median	mean ± SD	p-value	median	mean ± SD	p-value	median	mean ± SD	p-value	median
CNN – ann 1	0.92 ± 0.05	.56	0.93	2.10 ± 3.34	.83	1.00	0.48 ± 0.67	.67	0.35	10.2 ± 9.1	.28	7.1
CNN – ann 2	0.91 ± 0.05	.69	0.93	2.08 ± 3.41	.94	1.07	0.50 ± 0.68	.96	0.35	9.69 ± 9.19	.57	7.72
ann 1 – ann 2	0.92 ± 0.04		0.93	1.20 ± 0.65		1.00	0.34 ± 0.19		0.36	6.93 ± 5.32		4.53

Note.—Dice index, Hausdorff distance, surface-to-surface distance (S2S), and relative volume error (RVE) of the model compared withannotator (ann) 1, annotator 2 and both annotators of the contrast-enhanced T1-weighted model. Results of the (a) validation set and (b) independent test set are shown.CNN – convolutional neural network. P-values denotes Wilcoxon signed ranks test between this quantitative score and corresponding score of annotator1-annotator 2 (the third row).

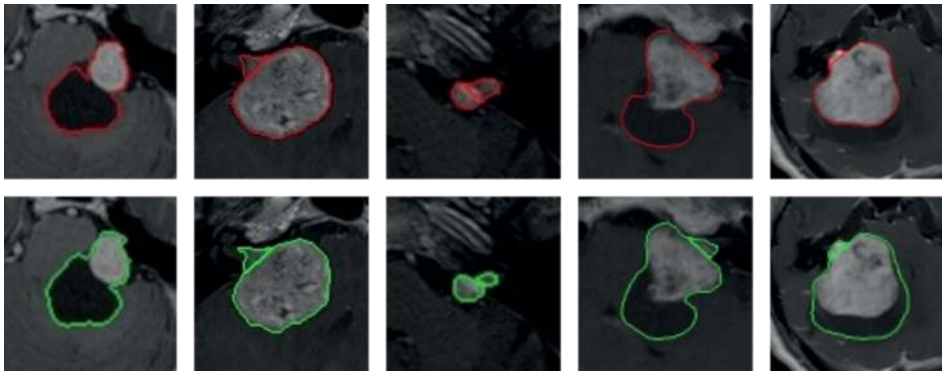


Figure 4. Examples of different (cystic, large, small) vestibular schwannoma whole tumor annotations, including the separation between the intra- and extrameatal tumor parts, of contrast-enhanced T1-weighted MRIs. The first row shows the convolutional neural network (CNN) predictions in red, and the second row shows the delineation of annotator 1 in green. The first, fourth and fifth tumors are potentially hard to delineate for the CNN due to the large peripheral cystic tumor parts. The Dice scores of these patients were 0.96, 0.96, 0.91, 0.93 and 0.72, respectively, and the surface-to-surface distances (mm) were 0.39, 0.21, 0.24, 0.35 and 3.44, respectively.

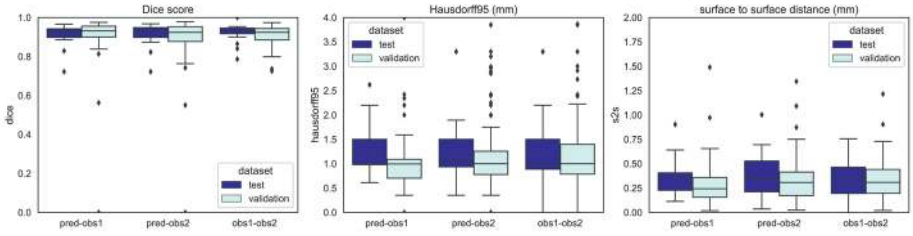


Figure 5. Quantitative measures of whole tumor convolutional neural network performance compared with the two annotators on contrast enhanced T1-weighted MRIs. Inter-annotator variability is also shown (obs 1-obs 2). From left to right the Dice indices, 95% Hausdorff distance (Hausdorff95) and surface-to-surface (S2S) distance boxplots are shown. The validation set results are shown in sky blue and the independent test set in dark blue.pred = CNN prediction, obs = observer.

3.4 Performance by Tumor Size

In the supplemental material (Fig. C) the results of the performance per size category are shown. Whole tumor results show a pattern of higher Dice indices for larger tumors, which was expected since the Dice index is sensitive to size. S2S was very similar in all size groups (<0.5mm), although S2Swere slightly larger in larger tumors ($p<0.001$). Results of intra- and extrameatal tumor parts show stable performance, except for four outliers in the small tumors (inaccurate extrameatal segmentation) and three outliers in giant tumors (false intrameatal tumor detection). In these tumors, there were some differences between model and human delineation for a completely intrameatal tumor with or without a tiny extrameatal part (small) or an extrameatal tumor with or without an intrameatal part (giant).

3.5 Outcomes of Observer Study

Agreement between the two observers before the consensus meeting on whole tumor segmentation quality was 131/134 (98%) for the human annotators and 127/134 (95%) for the CNN.

CNN segmentations of the whole tumor were considered comparable to the human segmentations in 103/111 (93%) of cases in the validation set and 20/23 (87%) in the test set. The CNN segmentations were rated better than the human segmentations in 2/111 (2%) and 2/23 (9%) of cases in the two datasets, respectively. Intrameatal segmentations were rated comparable to or better than human segmentations in 100/106 (94%) and 22/23 (96%) in the validation and test sets, respectively. For extrameatal segmentations, these percentages were 83/97 (86%) and 18/22 (82%).

In addition, the observers considered 104/111 (94%, validation set) and 20/23 (87%, test set) of whole tumor CNN segmentations satisfactory. Intrameatal tumor parts were considered satisfactory in 100/104 (94%, validation set) and 22/23 (96%, test set) of segmentations. Extrameatal tumor parts were considered satisfactory in 90/97 (93%, validation set) and 18/22 (82%) (test set) of segmentations. For human segmentations of the intrameatal tumor, 98/104 (94%) in the validation and 23/23 (100%) in the test set were rated satisfactory. Other satisfaction levels of the human segmentations were 110/111 (99%,validation set) and 22/23 (96%) (test set) for the whole tumor and 89/97 (92%, validation set) and 21/22 (95%, test set) for the extrameatal tumor part.

4. DISCUSSION

To our knowledge, this is the first study which presents the results of a multicenter, multivendor automated vestibular schwannoma segmentation tool. The developed 3D CNN-tool measured tumor volume with very high accuracy on contrast-enhanced T1-weighted MRIs and T2-weighted MRIs. The S2S distances were between 0.4 and 0.9 mm, which was lower than the median slice thickness of 1.0 mm. The observer study suggests that the tool performs comparably to human delineation in 87-93% of the cases.

The contrast-enhanced T1-weighted MRI model provided excellent S2S distances and Dice indices. However, the standard deviations of the Hausdorff distances were remarkably large in the test set due to two outliers which contained peripheral cysts in the extrameatal part. The model did have difficulties with tumors containing large peripheral cysts (see supplemental material Fig. A for examples), which were sometimes partially included by the model.

Evaluation of the model on the publicly available dataset of Shapey et al. showed robust performance on contrast-enhanced T1-weighted images.[15] Interestingly, the ground-truth delineations of Shapey et al. are smaller than those used in the current study, as shown in supplemental Figure D, reducing Dice index from 0.93 to 0.88.[7] When erosion (3x3 kernel) was performed on model delineation, Dice index improved again to 0.93±0.03, supporting this observation. The delineations by Shapey et al. were used for radiotherapy purposes, where preventing damage to the surrounding tissue is important, warranting conservative delineation. We did not compare the T2-weighted images of the publicly available dataset to those in our dataset given differences in the imaging characteristics (echo time and repetition time) and region of interest (whole brain vs. cerebellopontine angle region).

In our study, CNN performance on T2-weighted MRI was slightly less accurate with more uncertainty compared with the contrast-enhanced T1-weighted images. This was particularly the case in polycystic tumors, where the tumor border was hard to distinguish from the cerebrospinal fluid solely on T2 (supplemental fig. B). In one case, the model could not distinguish a small tumor obliterating the internal meatus. In another single case, the model detected the contralateral eye as a false positive volume outside the region of interest.

The RVE values of the whole tumor ranged from 8-12%, compared to 9-10% inter-annotator volume differences. Only the T2 model in the validation set had a larger RVE of 25%.The performance of our CNN compared with human volume measurement is below previously reported inter-annotator variabilities ranging from 15-20%[3-5], and also below the generally accepted threshold of 20% before volume increase is considered growth. Two dimensional measurements are advised in the consensus guidelines but have high intra-observer variabilities ranging from 10-40%.[2-5] Volume measurement is more accurate, and the proposed tool can reduce the workload which has been a barrier for clinical adoption, enabling the shift from 2D measurement. Since documented detection and evaluation of tumor growth is one of the main factors that indicate the need for treatment, be it surgical removal or irradiation, this is of notable clinical relevance.

A unique attribute in vestibular schwannoma research is the integration of an observer study. Determining a ground truth is necessary in artificial intelligence imaging studies. The reliability of the ground truth is uncertain when human observer performance is suboptimal, as described above. Our observer study allowed evaluation of the comparability between CNN segmentation and human segmentation, the reference standard. Our results showed that the CNN tool performs comparably to human observers in the vast majority of cases, supporting the quantitative results that the tool is feasible and

robust for usage in clinical practice. Whole tumor delineations performed slightly better than the extrameatal delineations, which should be considered when using the tool in clinical practice as extrameatal tumor progression is of particular interest for treatment decisions.

Artificial intelligence tools for vestibular schwannoma segmentation that have been previously proposed were all performed on data from a single center.[5,6,7] In clinical practice, however, diagnostic and follow-up scans are often performed in different centers using a variety of scanners and MRI protocols. In addition to its documented performance in a multicenter, multivendor setting, our method contains three features that make the tool more suitable for clinical practice compared to previous automated vestibular schwannoma delineation methods. First, the tool can distinguish between the intra- and extrameatal parts of the tumor. This distinction is important for clinical decision-making, as extension and progression of the extrameatal part usually determines the need for intervention. For this reason, current tumor staging systems are based mainly on the extrameatal dimensions of the tumor, while the intrameatal part is not measured.[2, 16] Second, the proposed tool can also delineate on solely T2-weighted MRI. Given the ongoing debate on use of gadolinium, this is a valuable feature.[17] Third, unlike previous models, our network is a fully 3D network that enables complete use of intra-slice information.

This study has some inherent limitations. First, the study was performed using retrospective MRI data. While this is an accepted method for the development of a new tool, some bias may be introduced by using older MRIs with suboptimal image quality and resolution. Therefore, accuracy and efficacy should also be investigated in prospective studies before clinical implementation and use. Second, for training of the T2 model, the registered human T1 delineations were used. This might have resulted in a suboptimal ground truth for the T2 model, although the reported tumor size correlations between T1 and high-resolution T2 were high.[18, 19] Third, the model is only trained on data before treatment and cannot be used for follow-up after surgery or radiotherapy without retraining.

Implementation of the CNN tool in clinical practice could lead to more accurate volume measurements of vestibular schwannoma at diagnosis and during follow-up, while reducing the workload of radiologists. Tumor volume change over time is a decisive factor in clinical decision making, and future research should focus on the tool's performance in a prospective study and its impact on clinical practice. The tool might be improved using post processing to reduce the false positive volumes outside the region of interest. In addition, the algorithm used for development of the tool could be adapted to analyze

other slow-growing skull base pathologies that are typically approached by a wait and scan policy, such as meningiomas.[20]

The proposed CNN model delineated vestibular schwannoma from MRI with excellent accuracy, comparable to human performance in the majority of cases. The CNN tool made the clinically relevant distinction between intra- and extrameatal tumor parts. The study shows the feasibility of automatically detecting and evaluating vestibular schwannoma with or without contrast administration in large datasets acquired from multiple medical centers and MRI vendors.

REFERENCES

1. Carlson ML, Link MJ. Vestibular Schwannomas. N Engl J Med 2021;384(14):1335-1348. doi: 10.1056/nejmra2020394
2. Kanzaki J, Tos M, Sanna M, Moffat DA, Monsell EM, Berliner KI. New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma. Otol Neurotol 2003;24(4):642-648; discussion 648-649. doi: 10.1097/00129492-200307000-00019
3. Varughese JK, Wentzel-Larsen T, Vassbotn F, Moen G, Lund-Johansen M. Analysis of vestibular schwannoma size in multiple dimensions: a comparative cohort study of different measurement techniques. Clin Otolaryngol 2010;35(2):97-103. doi: 10.1111/j.1749-4486.2010.02099.x
4. Mackeith S, Das T, Graves M, Patterson A, Donnelly N, Mannion R, Axon P, Tysome J. A comparison of semi-automated volumetric vs linear measurement of small vestibular schwannomas. Eur Arch Otorhinolaryngol 2018;275(4):867-874. doi: 10.1007/s00405-018-4865-z
5. van de Langenberg R, de Bondt BJ, Nelemans PJ, Baumert BG, Stokroos RJ. Follow-up assessment of vestibular schwannomas: volume quantification versus two-dimensional measurements. Neuroradiology 2009;51(8):517-524. doi: 10.1007/s00234-009-0529-4
6. Lees KA, Tombers NM, Link MJ, Driscoll CL, Neff BA, Van Gompel JJ, Lane JI, Lohse CM, Carlson ML. Natural History of Sporadic Vestibular Schwannoma: A Volumetric Study of Tumor Growth. Otolaryngology–Head and Neck Surgery 2018;159(3):535-542. doi: 10.1177/0194599818770413
7. Shapey J, Wang G, Dorent R, Dimitriadis A, Li W, Paddick I, Kitchen N, Bisdas S, Saeed SR, Ourselin S, Bradford R, Vercauteren T. An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI. J Neurosurg 2019:1-9. doi: 10.3171/2019.9.JNS191949
8. Lee C-C, Lee W-K, Wu C-C, Lu C-F, Yang H-C, Chen Y-W, Chung W-Y, Hu Y-S, Wu H-M, Wu Y-T, Guo W-Y. Applying artificial intelligence to longitudinal imaging analysis of vestibular schwannoma following radiosurgery. Sci Rep 2021;11(1). doi: 10.1038/s41598-021-82665-8
9. George-Jones NA, Wang K, Wang J, Hunter JB. Automated Detection of Vestibular Schwannoma Growth Using a Two-Dimensional U-Net Convolutional Neural Network. The Laryngoscope 2021;131(2). doi: 10.1002/lary.28695
10. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 2021;18(2):203-211. doi: 10.1038/s41592-020-01008-z
11. Isensee F, Petersen J, Klein A, Zimmerer D, Paul, Kohl S, Wasserthal J, Gregor, Wirkert S, Klaus. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. arXiv pre-print server 2018. doi: None arxiv:1809.10486
12. Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2010;29(1):196-205. doi: 10.1109/tmi.2009.2035616
13. Shamonin DP, Bron EE, Lelieveldt BP, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front Neuroinform 2013;7:50. doi: 10.3389/fninf.2013.00050
14. Rai R, Holloway LC, Brink C, Field M, Christiansen RL, Sun Y, Barton MB, Liney GP. Multicenter evaluation of MRI-based radiomic features: A phantom study. Med Phys 2020;47(7):3054-3063. doi: 10.1002/mp.14173
15. Shapey J, Kujawa A, Dorent R, Wang G, Bisdas S, Dimitriadis A, Grishchuck D, Paddick I, Kitchen N, Bradford R, Saeed S, Ourselin S, Vercauteren T. Segmentation of Vestibular Schwannoma from

Magnetic Resonance Imaging: An Open Annotated Dataset and Baseline Algorithm [Data set]. The Cancer Imaging Archive2021.

16. Koos WT, Day JD, Matula C, Levy DI. Neurotopographic considerations in the microsurgical treatment of small acoustic neurinomas. J Neurosurg 1998;88(3):506-512. doi: 10.3171/jns.1998.88.3.0506
17. Buch K, Juliano A, Stankovic KM, Curtin HD, Cunnane MB. Noncontrast vestibular schwannoma surveillance imaging including an MR cisternographic sequence: is there a need for postcontrast imaging? J Neurosurg 2019;131(2):549-554. doi: 10.3171/2018.3.jns1866
18. Tolisano AM, Wick CC, Hunter JB. Comparing Linear and Volumetric Vestibular Schwannoma Measurements Between T1 and T2 Magnetic Resonance Imaging Sequences. Otol Neurotol 2019;40(5S):S67-S71. doi: 10.1097/mao.0000000000002208
19. Pizzini FB, Sarno A, Galazzo IB, Fiorino F, Aragno AMR, Ciceri E, Ghimenton C, Mansueto G. Usefulness of High Resolution T2-Weighted Images in the Evaluation and Surveillance of Vestibular Schwannomas? Is Gadolinium Needed? Otol Neurotol 2020;41(1):e103-e110. doi: 10.1097/mao.0000000000002436
20. Whittle IR, Smith C, Navoo P, Collie D. Meningiomas. The Lancet 2004;363(9420):1535-1543. doi: 10.1016/s0140-6736(04)16153-9

Automated 2-Dimensional Measurement of Vestibular Schwannoma: Validity and Accuracy of an Artificial Intelligence Algorithm

Olaf Neve
Stephan Romeijn
Yunjie Chen
Larissa Nagtegaal
Willem Grootjans
Jeroen Jansen
Marius Staring
Berit Verbist
Erik Hensen

Otolaryngol Head Neck Surg. 2023 Dec;169(6):1582-1589
doi: 10.1002/ohn.470

ABSTRACT

Objectives

Validation of automated two-dimensional (2D) diameter measurements of vestibular schwannomas on MRI.

Study design

Retrospective validation study using two datasets containing MRIs of vestibular schwannoma patients.

Setting

University hospital in the Netherlands

Methods

Two datasets were used, one containing one scan per patient (n= 134) and the other containing at least three consecutive MRIs of 51 patients, all with contrast-enhanced T1 or high-resolution T2 sequences. 2D measurements of the maximal extrameatal diameters in the axial plane were automatically derived from a 3D-convolutional neural network compared to manual measurements by two human observers. Intra- and interobserver variabilities were calculated using the intraclass correlation coefficient (ICC), agreement on tumor progression using Cohen's kappa.

Results

The human intra- and interobserver variability showed high correlation (ICC 0.98- 0.99) and limits of agreement of 1.7-2.1mm. Comparing the automated to human measurements resulted in ICC of 0.98 (95%CI 0.974;0.987) and 0.97 (95%CI 0.968;0.984), with limits of agreement of 2.2 and 2.1 mm for diameters parallel and perpendicular to the posterior side of the temporal bone, respectively. There was satisfactory agreement on tumor progression between automated measurements and human observers (Cohen's kappa 0.77), better than the agreement between the human observers (Cohen's kappa 0.74).

Conclusion

Automated 2D diameter measurements and growth detection of vestibular schwannomas are at least as accurate as human 2D measurements. In clinical practice, measurements of the maximal extrameatal tumor (2D) diameters of vestibular schwannomas provide important complementary information to total tumor volume (3D) measurements. Combining both in an automated measurement algorithm facilitates clinical adoption.

INTRODUCTION

Vestibular schwannomas are benign intracranial tumors arising from the eighth cranial nerve. Patients typically present with audiovestibular symptoms such as hearing loss, balance problems and/or tinnitus. Other symptoms include headache, facial paresis or numbness.[1-3] A small majority of vestibular schwannomas are non-progressive, justifying active surveillance, with regular MRI as the preferred management strategy.[4] However, some tumors are progressive, which ultimately can lead to brainstem compression or intracranial hypertension. To prevent these potentially life-threatening conditions, progressive tumors are usually treated with either radiotherapy or surgery.

The accurate assessment of tumor progression is essential in clinical decision-making. Currently, tumor progression is determined based on the manual diameter measurements of subsequent MRIs.[5] However, these measurements have considerable errors, with reported intra- and interobserver variabilities ranging between 10% and 40%.[6-8] Compared to diameter measurements, volume measurements are considered to be more reliable for the detection of growth, however, these measurements are time-consuming.[6, 8, 9] For that reason, volume measurements have not widely been adopted in clinical practice yet, neither manual nor by semi-automated volume measurement algorithms.[7]

To overcome this problem, several fully automated volume measurement algorithms have been developed.[10-13] These algorithms use deep learning techniques to determine tumor volume and show excellent performance compared to human volume measurements. The wider implementation of these algorithms has been hampered by the fact that they have been trained on single-center data, using single-vendor MR scanners with limited variation in scan protocol. Therefore, the performance of these algorithms in different clinical settings is less reliable and requires additional external validation. We have recently developed an algorithm for the automated measurement of vestibular schwannomas that is based on multivendor, multicenter MR data, that has been validated externally and is applicable to different MR sequences.[13]

In current clinical practice, treatment decisions as well as consensus-based classifications such as those proposed by Koos et al.[14] and Kanzaki et al. are not based on tumor volume but on extrameatal tumor diameters.[5] Treatment decisions and tumor classifications focus on the extrameatal tumor parts rather than whole tumor volume, because the extrameatal extension is the closest proxy measurement to the anatomical relation and impact of the tumor to critical adjacent structures such as the brain stem.[5] So, whereas volume change is superior in detecting tumor progression, extrameatal

diameters provide essential additional information on the direction of tumor extension and progression. In 2018, a survey study showed that 91% of the members of the North American Skull base Society would observe a small tumor (<15mm cerebellar pontine angle (CPA)) until growth was detected.[15] Since then, several papers have been published arguing for observation in small but progressive tumors (CPA < 15 mm) and a size threshold for active treatment was introduced, based on extrameatal tumor diameters, emphasizing the complementary value of tumor diameters to tumor volume measurements.[16, 17]

Therefore, this study aimed to validate an algorithm to measure extra-meatal tumor diameters as an addition to a previously reported automated volume measurement algorithm.[13] Combining automated two-dimensional(2D) and tumor volume (3D) measurements in one algorithm would result in a robust tool suited to support treatment decisions in current clinical practice.

METHOD

This retrospective study was performed in a university hospital in the Netherlands, an expert center for vestibular schwannoma. The protocol has been reviewed by the Medical Research Ethics Committee Leiden Den Haag Delft (G19.115), which granted an exemption for informed consent.

Measurement algorithm

This study aimed to extend the existing in-house developed automated volume measurement model with automated 2D measurements, i.e. the maximal extrameatal tumor diameters in the axial plane. To do so, the automated 2D measurements were compared with repeated human measurements of two observers (ON and SR). The intra- and interobserver variability were analyzed. Second, the mean diameter of the two observers was used as ground truth to evaluate the automated measurements. All diameters were measured according to the consensus guidelines as proposed by Kanzaki et al.[5], i.e. the largest extrameatal diameter parallel to the petrous bone was measured first, followed by the largest extrameatal diameter perpendicular to the line drawn to acquire the first diameter (i.e., perpedicular to medial surface of the petrous bone).

The automated volume measurement algorithm, based on a convolutional neural network (CNN), was previously developed and validated by our research group[13] using the nnU-net framework.[18] For vestibular schwannomas, we used a 3D U-Net with five encoder and decoder layers, detailed in a previous publication by Neve et al.[13] The

model was trained and validated on scans from 37 different centers and was able to delineate tumors on contrast-enhanced T1 and on high-resolution (hr) T2.[13] Furthermore, the performance was externally validated on the publicly available dataset by Shapey et al.[13, 19] In addition, the model was able to differentiate between the intra- and extrameatal tumor parts.

For the automated 2D measurements, the border between intra- and extrameatal tumor segmentations was used to select the plane parallel to the petrous bone, and orthogonal to the axial plane to mimic the clinical procedure. Using this plane, the largest parallel diameter was chosen from all axial slices in the segmentation. Consecutively, the largest diameter perpendicular to the parallel plane was derived from the same slice.

Design

Three analyses were performed. First, the intra- and interobserver variability of human 2D measurements was evaluated. Second, the accuracy of the automated 2D measurement was evaluated by comparing them to the human 2D diameters. Third, the capability to detect tumor progression on consecutive scans based on automated 2D diameters was evaluated.

Study population

Two different datasets were used in this study. The first was used for the development of the automated segmentations from the study by Neve et al. ('development dataset'). This development dataset contained 134 patients with one contrast-enhanced T1-weighted MRI. Of all MRIs the diameters were measured by two human observers (ON and SR) and in a subset of 50 patients both observers measured the diameters twice to assess the intraobserver variability.

Second, we randomly selected a data of 51 patients from vestibular schwannoma patients at our center, that had not been part of the first dataset. These 51 patients had at least three consecutive MRIs without intercurrent active treatment (surgery or radiotherapy). This dataset (the 'longitudinal dataset') was used to assess tumor progression. Both observers (ON and SR) measured the diameters of all MRIs. In challenging cases the observers consulted a senior head and neck radiologist (BV) with 22 years of experience to discuss the right plane and measurement. This consultation was performed in 6% of the MRIs. When contrast-enhanced T1 was not acquired, the measurement was performed on hrT2. Using both T1 and hrT2 mimics the clinical setting in which either one or both sequences are used in follow-up.

For the evaluation of the intra- and interobserver variability of the human 2D measurements and the accuracy of the automated 2D measurements, both the development and longitudinal datasets were merged. Tumor progression analysis was performed on the longitudinal dataset, as this contained multiple consecutive scans per patient.

Statistical analysis

All analyses were performed in R version 4.1.1 using R-studio 1.4.1717 (Rstudio, PBC, Boston). The intra- and interobserver variability of human 2D measurements were evaluated by calculating the interclass correlation coefficient (ICC) and plotting Bland-Altman plots, containing the difference in measurement on the Y-axis and the mean of the measurements on the X-axis.20 Bland-Altman limits of agreement were calculated by the mean difference between the measurements ±1.96 times the standard deviation of the difference between measurements. CNN diameters were compared to the mean of the two human diameters to reduce the impact of human interobserver variability. Automated diameter outliers which exceeded the limits of agreement were analyzed by a senior head and neck radiologist (BV) and are discussed in the discussion section.

Longitudinal tumor progression was based on a cut-off value of ≥2mm difference between two consecutive scans. The mean of the two human measurements was used as ground truth. CNN diameter progression performance was evaluated using sensitivity, specificity, and accuracy. In addition, Cohen kappa was calculated. These results were compared to the agreement on tumor progression between the two human observers. The correlation of the maximal diameter in the axial plane (parallel or perpendicular) with the maximal diameter of the entire 3D extrameatal component was evaluated using the ICC.

RESULTS

Patient characteristics of both datasets are shown in Table 1 and technical characteristics in Table 2. In the longitudinal dataset 9 out of 153 scans could not be extracted from the picture archiving and communication system due to technical incompatibilities. The tumor size and cystic component distributions differ between the datasets. Patients in the first dataset, used for the development of the automated volume CNN, were selected to have a large variety of tumor sizes. In contrast, the longitudinal dataset was a random sample of all patients treated at our center. These selection methods might explain the difference in patient age since patients with larger tumors tend to be younger than patients with smaller tumors. Examples of the automated diameters are shown in Fig. 1.

Table 1. Patient characteristics

	Development dataset	Longitudinal dataset
N	134	51
MRI scans per patient	1	3
Age in years (sd)	53.5 (12.0)	61 (10.4)
Sex male	64 (48%)	28 (55%)
Cystic component	63 (47%)	7 (14%)
Tumor size		
intrameatal	28 (21%)	20 (39%)
small (0-10mm)	19 (14%)	18 (35%)
medium (11-20mm)	26 (19%)	11 (22%)
moderately large (21-30mm)	24 (18%)	1 (2%)
large (31-40mm)	24 (18%)	1 (2%)
giant (>40mm)	13 (10%)	0

Table 2. Technical characteristics

	Development dataset	Longitudinal dataset	
	Contrast-enhanced T1-weighted MRI	Contrast-enhanced T1-weighted MRI	T2-weighted MRI
No. of scans	134	116	28
In-plane resolution	0.35x0.35 (0.27x0.27 - 1.0x1.0)	0.5x0.5 (0.27x0.27 – 1.13x.1.13)	0.35x0.35 (0.20x0.20 – 0.55x0.55)
In-plane matrix	400x400 (256x208 - 560x560)	352x352 (256x192 – 640x520)	512x512 (256x256 - 1024x1024)
TE(msec)	9 (2.38 - 20)	8.9 (2.38-22)	176.141 (1.968-263)
TR(msec)	602.10 (8.76 - 2200)	450 (6.84-1900)	1200 (5.42-5110)
Section thickness	1.0 (0.9 - 5.0)	2 (0.6-6.0)	1 (0.5-3)

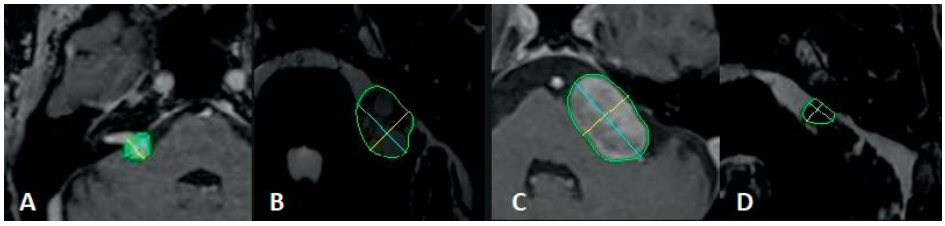


Figure 1.
Automated diameter measurements on contrast-enhanced T1 (a, c) and hrT2 (b, d) MRI. Automated tumor segmentations (green line), largest extrameatal diameters parallel (blue line) and perpendicular (yellow line) to the petrous bone.

Intra- and interobserver variability

Interobserver differences of the human 2D measurements are shown in Fig. 2A and B. The ICCs of the parallel and perpendicular measurements were both 0.984 (95% confidence interval (CI) 0.976;0.989), however the limits of agreement were 1.7 and 1.9 mm, respectively.

Intraobserver differences provided similar ICCs for parallel (0.995 95%CI 0.992;0.997) and perpendicular (0.989 95%CI 0.981;0.993) measurements, and the limits of agreement were 1.9 and 2.1 mm, respectively (shown in Fig. 2C and D).

Automated 2D measurement

The correlation between human and CNN diameters was excellent, with ICCs of 0.98 (95%CI 0.974;0.987) and 0.97 (95%CI 0.968;0.984) for the parallel and perpendicular diameters, respectively. As is shown in Fig. 3, the model diameters were, on average, slightly larger than the human diameters, resulting in a mean difference between human and CNN of 0.7 mm for parallel and 0.8 mm for perpendicular measurements. The limits of agreement were 2.2 mm for the parallel diameter and 2.1 for the perpendicular diameter.

Next, as the model is not confined to measurements in the axial plane, we evaluated the correlation of the maximal diameter in the axial plane (parallel or perpendicular) with the maximal diameter of the entire 3D extrameatal component. We found an excellent

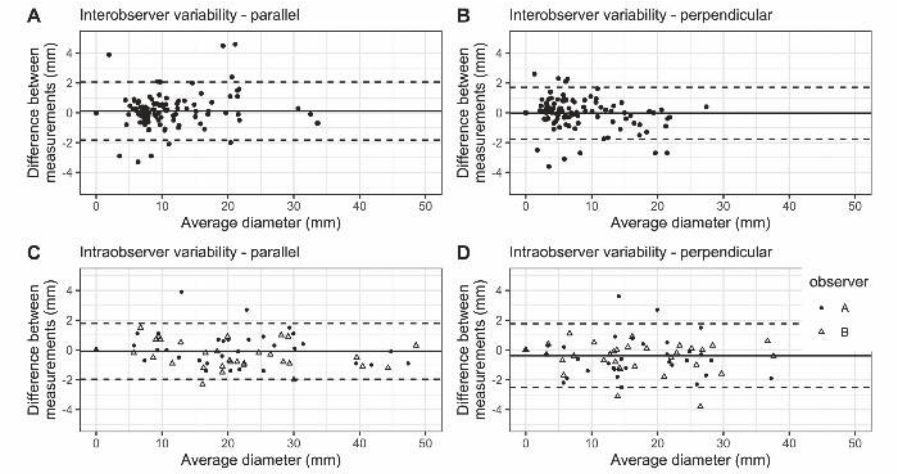


Figure 2.
Bland-Altman plots of intra- and interobserver variability of human-derived diameter measurements (A-D). Limits of agreement (dotted line). The mean difference between measurements (black line)

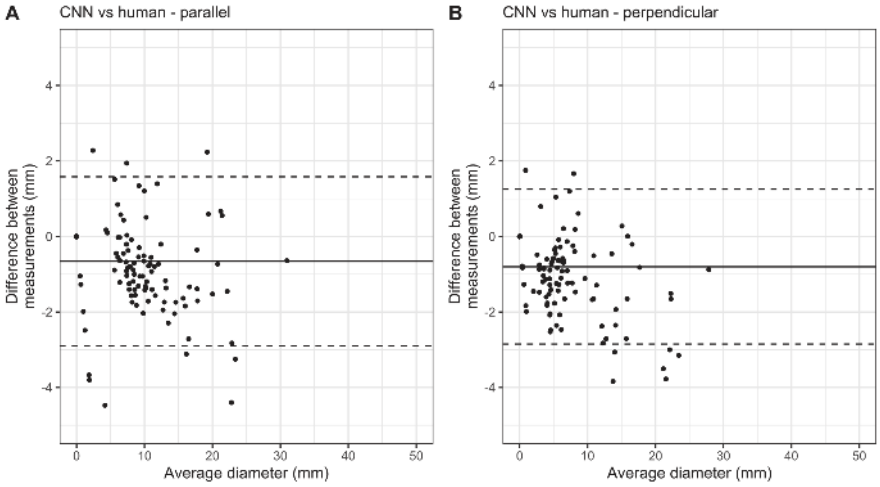


Figure 3.
Bland-Altmann plots of convolutional neural network (CNN) derived versus mean human-derived diameters (A-B). Limits of agreement (dotted line). The mean difference between measurements (black line)

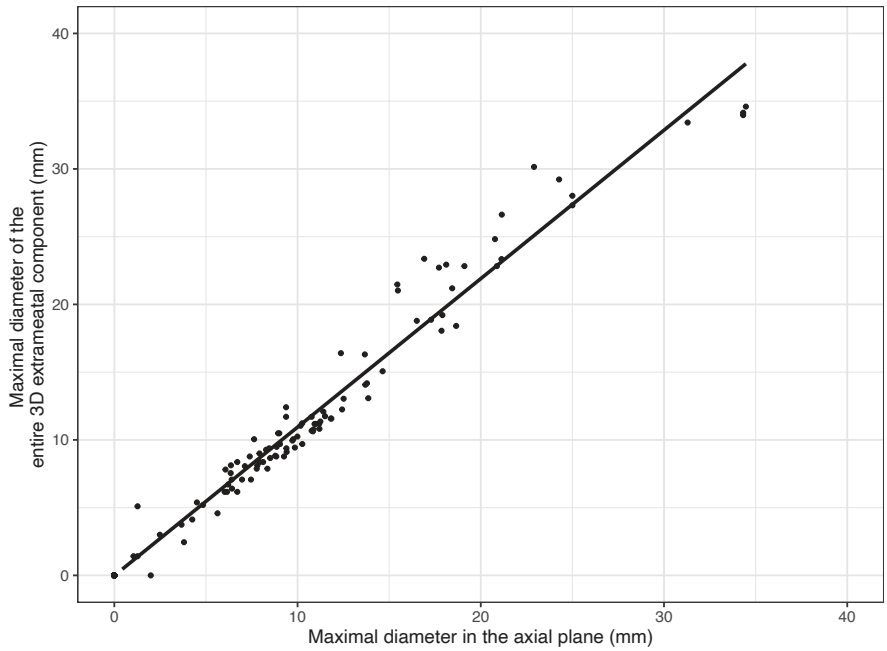


Figure 4.
Correlation between the maximal extrameatal diameter in the axial plane with the maximal diameter of the entire 3D extrameatal component.

ICC 0.974 (95%CI 0.970;0.984) between the largest diameter in axial plane and the largest diameter in the entire 3D extrameatal component, as shown in Fig. 4.

Tumor progression

Table 3 shows the evaluation of agreement on the diameter progression of the CNN compared to the human measurements and agreement on the diameter progression of the two human observers. The agreement on tumor progression between the CNN and the mean of the two human observers resulted in a Cohen's kappa of 0.77, indicating substantial agreement. Cohen's kappa of the agreement between the two human observers was 0.74. Also, the sensitivity, specificity, and accuracy of the CNN compared to the mean of the two human observers were comparable to these values when comparing the two human observers.

DISCUSSION

To our knowledge, this is the first study to propose an automated vestibular schwannoma 2D measurement algorithm using artificial intelligence techniques. The current study shows an intra- and interobserver measurement error of 1.7- 2.1mm in the 2D diameter measurement of vestibular schwannomas. The automated measurements were comparable to human measurements. The automated algorithm was able to detect tumor progression on consecutive MRI using either contrast enhanced T1 or hrT2 sequences.

On average, the automated measurements were 0.7-0.8 mm larger than the human measurements. This difference may in part be caused by the fact humans decide by eyeballing what would be the maximal line to measure the diameter, while the automated method really maximizes this mathematically based on contrast differences. In addition, automated segmentations use contrast differences and maximize the segmentation on pixel level by including the contour lines of the tumor. Indeed, further analysis of outliers revealed that automatic measurements included the entire thickness of the segmentation contour line. Another explanation for the outliers was the difference between the algorithm and human observers in separating the intra- and extrameatal tumor parts. When a larger proportion of tumors is considered extremeatal, this affects the extrameatal diameters. The segmentation algorithm is trained on human segmentations of the whole tumor and the intra- and extrameatal tumor parts. The algorithm is not trained to detect specific anatomical structures such as the edge of the petrous bone, to determine the difference between intra- and extrameatal tumor parts. However, the use of other anatomical structures is incorporated indirectly since the human

observers who annotated the training set did make use of the surrounding anatomical structures to determine the difference between the intra- and extrameatal tumor parts.

Both the intra- and interobserver variability of diameter measurements in vestibular schwannomas in the current study (respectively 0.98 and 0.99) are similar to previously reported ICCs. Langenberg et al.[8] and MacKeith et al.[7] have reported an ICC of 0.95 for interobserver agreement on diameter. Tolisano et al. have reported a similar ICC of 0.98 and 0.99 for interobserver agreement on contrast-enhanced T1 and hrT2 sequences. The intraobserver variability has previously been described by MacKeith et al.[7] and Coelho et al.[21] ranging from 0.92-0.98. The ICC of the automated measurements compared to the mean of the two human measurements is similar with 0.98 and 0.97, indicating that the automated measurement is acceptable for use in clinical practice.

The study by Hougaard et al.[22] also used Bland-Altman limits of agreement for 2D measurement. They have reported limits of agreement for interobserver variability of 2.8 mm for parallel and 2.2 mm for perpendicular diameters. The intraobserver limits of agreement were smaller (2.6 mm and 1.9 mm). In the current study the differences between interobserver (1.7 mm and 1.9 mm) and intraobserver (1.9 mm and 2.1 mm) limits of agreement were smaller and the interobserver limits were even lower compared to Hougaard et al. Considering the amount of variability in human diameter measurement, the performance of the automated diameter measurements (2.2 mm and 2.1 mm) is within the limits of human measurements.

The agreement on tumor progression based on diameter measurements on consecutive MRIs have been analyzed by Tolisano et al. using Cohen's kappa. They reported a Cohen's kappa of 0.56 and 0.61 for contrast-enhanced T1 and hrT2 sequences, respectively.[23] These agreement measures are slightly lower compared to Cohen's kappa (0.74) found in the current study when the agreement between two human observers was compared. Automated diameter measurements (0.77) even outperformed this, showing the capabilities of the CNN to detect tumor growth.

This study has some limitations. As this analysis was performed on retrospective data, reliability needs to be validated using prospective data before use in clinical practice. In addition, the dataset contained a small number of cystic tumors. These tumors are more challenging to delineate and could be prone for less accurate automated measurements. However, this is also true for manual measurements. Automated recognition of these cystic tumors could be a valuable improvement to the model as this could be used to alert radiologists to manually check the measurement of these tumors, thereby facilitating the clinical adoption of the tool. Furthermore, the dataset also contained

intrameatal tumors. Although this reflects clinical practice, the inclusion of intrameatal tumors was suboptimal for the validation of the automated extrameatal diameter measurements. Furthermore, the plane of the parallel extrameatal diameters was based on the border between intra- and extrameatal tumor part. As a consequence the algorithm was unable to measure diameters of completely extrameatal tumors. In contrast, completely intrameatal tumors were detected and categorized as an extrameatal diameter of 0 mm.

Tumor diameter measurements show wide intra- and interobserver variability. Tumor volume measurements are widely accepted to more reliably detect tumor progression.[8] However, volumetric measurements hold limited information about the direction of tumor extension. Furthermore, current consensus classifications systems, such as proposed by Kanzaki et al. and Koos et al., are based on (extrameatal) diameter measurements. As the direction of the volumetric tumor progression is essential information in clinical decision-making, extrameatal diameters provide important information complementary to tumor volume (change). By including both measures in a reliable automated system that is able to deal with both contrast enhanced T1 and hrT2 weighted MR imaging, we aim to provide a robust algorithm to support clinical decision-making in vestibular schwannoma patients.

The current algorithm is able to measure tumor diameters and volumes efficiently and consistently, which can be of added value in clinical practice compared to the currently used manual measurement limited to 2D diameters. Automated, consistent measurement of both diameters and volumes in consecutive scans could improve the accuracy of tumor growth detection as well as provide therapy-relevant information, while saving time and costs. It could therefore be a useful and efficient tool for multicenter vestibular schwannoma research and care, however future research is needed to evaluate the impact of incorporating automated tumor measurements and progression detection on clinical practice.

CONCLUSION

The accuracy of automated 2D measurements is comparable to manual 2D diameter measurements. Adding 2D diameters to tumor 3D volume measurements in one automated model provides a robust algorithm that can assist in clinical decision-making in vestibular schwannoma patients. The algorithm proposed in this study is able to deal with both contrast enhanced T1 and hrT2 weighted MR imaging of different MR scanner types and protocols, enabling its use in a multicenter setting.

REFERENCES

1. Arthurs BJ, Fairbanks RK, Demakas JJ, et al. A review of treatment modalities for vestibular schwannoma. Neurosurg Rev. 2011;34:265-279. doi:10.1007/s10143-011-0307-8
2. Management of Sporadic Vestibular Schwannoma, 48 407-422 (2015).
3. Matthies C, Samii M. Management of 1000 vestibular schwannomas (acoustic neuromas): clinical presentation. Neurosurgery. 1997;40:1-9; discussion 9-10.
4. Møller MN, Hansen S, Miyazaki H, Stangerup S, Caye-Thomasen P. Active Treatment is Not Indicated in the Majority of Patients Diagnosed with a Vestibular Schwannoma : A Review on the Natural History of Hearing and Tumor Growth. Current Otorhinolaryngology Reports. 2014;2:242-247. doi:10.1007/s40136-014-0064-7
5. Kanzaki J, Tos M, Sanna M, Moffat DA, Monsell EM, Berliner KI. New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma. Otol Neurotol. Jul 2003;24(4):642-8; discussion 648-9. doi:10.1097/00129492-200307000-00019
6. Varughese JK, Wentzel-Larsen T, Vassbotn F, Moen G, Lund-Johansen M. Analysis of vestibular schwannoma size in multiple dimensions: a comparative cohort study of different measurement techniques. Clin Otolaryngol. 2010;35(2):97-103. doi:10.1111/j.1749-4486.2010.02099.x
7. Mackeith S, Das T, Graves M, et al. A comparison of semi-automated volumetric vs linear measurement of small vestibular schwannomas. Eur Arch Otorhinolaryngol. 2018;275(4):867-874. doi:10.1007/s00405-018-4865-z
8. van de Langenberg R, de Bondt BJ, Nelemans PJ, Baumert BG, Stokroos RJ. Follow-up assessment of vestibular schwannomas: volume quantification versus two-dimensional measurements. Neuroradiology. Aug 2009;51(8):517-24. doi:10.1007/s00234-009-0529-4
9. Cross JJ, Baguley DM, Antoun NM, Moffat DA, Prevost AT. Reproducibility of volume measurements of vestibular schwannomas - a preliminary study. Clin Otolaryngol. 2006;31(2):123-129. doi:10.1111/j.1749-4486.2006.01161.x
10. Shapey J, Wang G, Dorent R, et al. An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI. J Neurosurg. Dec 6 2019:1-9. doi:10.3171/2019.9.JNS191949
11. Lee C-C, Lee W-K, Wu C-C, et al. Applying artificial intelligence to longitudinal imaging analysis of vestibular schwannoma following radiosurgery. Sci Rep. 2021;11(1)doi:10.1038/s41598-021-82665-8
12. George-Jones NA, Wang K, Wang J, Hunter JB. Automated Detection of Vestibular Schwannoma Growth Using a Two-Dimensional U-Net Convolutional Neural Network. The Laryngoscope. 2021-02-01 2021;131(2)doi:10.1002/lary.28695
13. Neve OM, Chen Y, Tao Q, et al. Fully Automated 3D Vestibular Schwannoma Segmentation with and without Gadolinium-based Contrast Material: A Multicenter, Multivendor Study. Radiology: Artificial Intelligence. 2022;4(4):e210300. doi:10.1148/ryai.210300
14. Koos WT, Day JD, Matula C, Levy DI. Neurotopographic considerations in the microsurgical treatment of small acoustic neurinomas. J Neurosurg. 1998;88(3):506-512. doi:10.3171/jns.1998.88.3.0506
15. Van Gompel J, Wiet R, Tombers N, et al. A Cross-sectional Survey of the North American Skull Base Society: Current Practice Patterns of Vestibular Schwannoma Evaluation and Management in North America. Journal of Neurological Surgery Part B: Skull Base. 2018;79(03):289-296. doi:10.1055/s-0037-1607319

16. Macielak RJ, Wallerius KP, Lawlor SK, et al. Defining clinically significant tumor size in vestibular schwannoma to inform timing of microsurgery during wait-and-scan management: moving beyond minimum detectable growth. J Neurosurg. 2021:1-9. doi:10.3171/2021.4.jns21465
17. Marinelli JP, Lohse CM, Carlson ML. Introducing an Evidence-Based Approach to Wait-And-Scan Management of Sporadic Vestibular Schwannoma. Otolaryngol Clin North Am. 2023;56(3):445-457. doi:10.1016/j.otc.2023.02.006
18. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods. 2021/02/01 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z
19. Shapey J, Kujawa A, Dorent R, et al. Data from: Segmentation of Vestibular Schwannoma from Magnetic Resonance Imaging: An Open Annotated Dataset and Baseline Algorithm [Dataset]. 2021. The Cancer Imaging Archive. doi:https://doi.org/10.7937/TCIA.9YTJ-5Q73
20. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. Feb 8 1986;1(8476):307-10.
21. Coelho DH, Tang Y, Suddarth B, Mamdani M. MRI surveillance of vestibular schwannomas without contrast enhancement: Clinical and economic evaluation. The Laryngoscope. 2018;128(1):202-209. doi:10.1002/lary.26589
22. Hougaard D, Norgaard A, Pedersen T, Bibby BM, Ovesen T. Is a redefinition of the growth criteria of vestibular schwannomas needed? Am J Otolaryngol. 2014;35(2):192-197. doi:10.1016/j.amjoto.2013.08.002
23. Tolisano AM, Wick CC, Hunter JB. Comparing Linear and Volumetric Vestibular Schwannoma Measurements Between T1 and T2 Magnetic Resonance Imaging Sequences. Otol Neurotol. 2019;40:S67-S71. doi:10.1097/mao.0000000000002208

CHAPTER 10

General Discussion

The aim of this thesis was to assess several aspects of VBHC in vestibular schwannoma care. In the first part, factors that influence the shared treatment decision and outcomes of management strategies relevant for patients, such as quality of life and employment have been assessed. In addition, the most effective method of acquiring the patients' feedback has been evaluated. In the second part of the thesis, the added value of data driven technologies in vestibular schwannoma care has been assessed. The application of two novel data driven technologies were evaluated: an automated analysis of the patients' experiences with the care delivery, and automated measurements of tumors (a key element in the treatment decision).

10.1 Relevant outcomes in vestibular schwannoma care

In **chapter 1**, four pillars of Dutch VBHC were described (Organizing care in an IPU, creating a core outcome set that is relevant to and measured in every patient, shared decision making, and an enabling technology platform). These pillars can empower patients to play a more prominent role in clinical decision-making but also in healthcare delivery.[1] In the field of vestibular schwannoma, the transition from disease or tumor-oriented healthcare towards patient-centered healthcare is still ongoing.

In the distant past, diagnostic and treatment options were limited and were associated with high mortality rates.[2] Impressive improvements in the last century in the fields of diagnosis and treatment have improved the patient's prognosis dramatically, and paved the way for the transition from a solely medical perspective on vestibular schwannoma care, with mostly clinical and tumor-oriented outcomes, towards a patient-centered perspective with a focus on the patients' functioning in daily life.

Increasing knowledge of the natural course of the disease, accompanied by a rising incidence and earlier detection of tumors, has changed the management of vestibular schwannoma. Active treatment is no longer the treatment option of choice at diagnosis in the majority of cases, as it was in the '90s.[3,4] Instead, active surveillance has become the strategy of first choice for most patients with small and indolent or slowly growing tumors.

With changing management strategies, attention to the quality of life of vestibular schwannoma patients has increased since 2000.[5] In 2011, a disease-specific vestibular schwannoma quality of life questionnaire was developed, which contributed to increased knowledge of the impact of the disease and its treatment on the quality of life of the patients.[6] Using this quality of life questionnaire as PROM in every vestibular schwannoma patient in an IPU seems therefore justified. Patient preferred outcomes such as hearing, balance, and tinnitus as reported by Pruijn et al are covered in the

disease-specific questionnaire.[7] These areas of interest were also reflected in the factors that influence patient decision making in **chapter 5**.

Chapter 2 showed that the long-term quality of life outcomes are stable over time and that there were no clinically relevant differences between the three treatment strategies. Furthermore, employment rates and absenteeism were similar across treatment strategies and at the same level as the age-matched Dutch population, as indicated in **chapter 3**. These findings provide insights into the ability of vestibular schwannoma patients to participate in society. Employment is essential not only from a financial perspective, but it also improves patients' self-worth and provides social connections.[8-10] The importance of employment for vestibular schwannoma patients was indeed emphasized in **chapter 5**, in which patients also considered the time to return to work after treatment as a factor in their treatment decision-making.

At the group level, the findings of **chapters 2 and 3** can help physicians to use, understand and interpret PROMs. Physicians need to have a reference when analyzing PROM outcomes of individual patients that have completed PROMs as part of the core outcome set. For laboratory results, reference values are commonplace and integrated with hospital information systems. Adequate use of PROMs during consultations requires knowledge of average PROM outcomes at a population level. Ideally, the averages can be matched on factors such as age, sex, and time since diagnosis, to compare individual patient results to a "patients-like-me" reference group. Additionally, minimal clinically important differences should be used both to assess changes over time within patients and to compare these to the reference level.[11,12]

The use of outcomes such as quality of life and employment at the individual level is more complex. The qualitative research in **chapter 5** showed that several factors influenced patients in their treatment decision process. Some identified factors were known medical or physician-based factors, such as tumor progression and treatment advice. But the qualitative study also showed new factors that might be less relevant from a medical or physician-based perspective. Examples of such factors were traveling time, holiday planning, and the ability to care for one's children after a specific treatment strategy (**chapter 5)**. In addition, the search for relevant information on the internet and experiences of relatives with treatment modalities for completely different diseases (e.g., radiotherapy for breast cancer) can inform and guide patient decision-making. All these factors, medical and non-medical, could be explored, deliberated, or addressed when a physician guides a patient toward shared decision-making. In addition, **chapter 5** showed that qualitative research, although it is rare in the vestibular schwannoma

literature, is complementary to quantitative research as it can answer different types of research questions.

The results of **chapters 2 and 3** provide awareness of and knowledge about outcomes relevant to patients. This may assist physicians in counseling patients through a shared decision-making process and aid patients in selecting the treatment most aligned with their life circumstances and context. **Chapters 2 and 3** showed no differences in long term quality of life outcomes and employment rates between the treatment strategies on a group level. This seemed surprising, since the patient groups per treatment strategy differ at baseline in tumor size, tumor growth and symptomology. It may indicate that current treatment decision making, which is primarily based on tumor characteristics and patient's comorbidity and preference, is effective in optimizing quality of life outcomes. It seems to justify that quality of life, although important, does not function as a criterium to categorically opt for or reject a specific treatment option in the decision-making process. These findings will therefore not change the current decision-making process in our center, where a conservative management strategy is the option of choice, and interventions are generally reserved for patients with large or progressive tumors or with specific symptoms. The patients' preferences weigh heavily when considering the different treatment options, especially when the success rate of the different options seems comparable.

In both using outcome data at the group level and incorporating outcome data into shared decision-making processes, it is imperative to acknowledge the potential risks associated with poor generalizability due to low response rates. As illustrated in **Chapter 4**, the method of delivery significantly influences response rates, leading to the underrepresentation of certain demographic groups. Recognizing these risks is essential for mitigating the potential biases that may arise during the interpretation of patient-reported outcome data.

10.2 Use of data driven technology for measuring patient experience

Measurement of PROMs as part of the core outcome set for vestibular schwannoma care contributes to VBHC, to quantify and monitor quality improvement. These PROMs, however, do not necessarily reflect the patients' experience of their care delivery. Patients' experiences are part of the quality assessment described by Donabedian **(chapter 1)** and are positively associated with other domains of quality of care, such as patient safety and effectiveness. [13-15]

Measuring experiences is performed in many service delivery industries and health care is no exception. Several PREM instruments are used to compare hospitals and contrib-

ute to continuous quality improvement. However, using the PREM results to improve healthcare delivery is often challenging. For example, the Netherlands Federation of University Medical Centers (NFU) yearly collects the same PREM data in all university hospitals. These results are used for benchmarking, but actual quality improvement initiatives in clinical practice based on these PREM outcomes are still scarce. [16] The use of PREMs often remains limited to that by managerial and supportive hospital staff. These findings align with results from the United Kingdom, where collecting PREMs is mandatory for hospitals, but the adoption of PREM-based quality improvement is falling short. [17]

Collecting patients' experiences without using the results is disrespectful to the patients who invest time and effort to complete the PREMs, and decreases the patients' willingness to participate in these surveys. Moreover, it is a waste of useful information that may contribute to better care delivery. Measuring for the sake of measuring or to comply with national registrations or legislation is time-consuming for healthcare providers and does not contribute to value creation. [18]

In **chapters 6 and 7,** open-ended PREM questions combined with automated computerized analysis of the answers provided valuable insights and actionable points for quality improvement. Especially the combination of quantitative and qualitative results offered helpful information to improve the vestibular schwannoma IPU. In-depth descriptions of experiences were classified and clustered automatically, reducing healthcare providers' workload. As described in **chapter 7**, the human component in the analysis was still essential to interpret the output, relate it to the local context of the vestibular schwannoma IPU, and translate it into measures to improve the care delivery.

Sometimes patients' experiences are seen as rather subjective or non-contributing to quality of care. [15, 19] This critique holds especially true for 'patient satisfaction', which is a judgment or rank of the experienced care and is influenced by a priori expectations. The measurement of 'patient experiences' focusses instead on what actually happened during the care delivery and thereby provides more objective information. [20] Moreover, patients' experiences should be seen as one of the modalities for assessing quality of care, not the only one. Patients' experiences can be complementary to clinical, process, and quality of life outcomes. They incorporate the patients' perspective into the assessment of quality of care and predominantly evaluate the first two quality domains described by Donabedian: structure and process. The patients' experiences can be seen as complementary in pursuing patient-centered care, and measuring them using PREMs can contribute to overall quality improvement and value creation.

10.3 Data-driven health care and workload

Data-driven technology can improve quality and reduce workload by automating tasks and analyzing the large amount of data acquired from every individual patient. Both aspects are essential in VBHC, in which value creation can be realized by improving outcomes or reducing costs. In this thesis, two newly developed applications have been presented that use data-driven technologies. Both applications can improve quality of data by improving the completeness or accuracy, while reducing the workload of clinicians using automated analysis of different types of data.

In **chapters 6 and 7,** the development and the added value of the AI-PREM were described. This tool provides automated open-text analysis, assisting clinicians in finding points for quality improvement based on the patient's experience. **Chapters 8 and 9** reported an automated measurement tool for vestibular schwannoma volume and extension on MRI. Using volume measurement instead of two-dimensional measurement is more accurate in detecting differences in tumor size, but time-consuming. [21] The automated tool can, therefore, improve measurement accuracy while reducing the workload of radiologists. (**chapter 8**) The increased accuracy helps to better detect growth (**chapter 9**), which is for clinicians but also for patients (**chapter 5**) one of the most important parameters in treatment decision making.

As mentioned in **chapter 1,** enabling technologies are essential to implement VBHC in daily clinical practice. Adoption of such applications has some ethical and practical consequences which should be considered and will be discussed in this section

First, artificial intelligence algorithms tend to be a black box without transparency about the underlying choices or mechanisms on which the outcome is based. [22] When looking at the outcome for a specific patient, it can be unclear whether relevant patient factors are in or excluded when using the outcomes of an algorithm. This lack of transparency hampers the ability for end users to assess the risk of errors. This is especially important when clinical decisions are based on the outcomes. [23] In both proposed tools in this thesis a human validation of the result is still necessary when using the tools in clinical practice. Second, the algorithms are as good or bad as the data they are trained on. Bias incorporated into the data collection can cause distorted results in groups with different baseline characteristics and may lead to unwanted bias or discriminatory outcomes. [22, 24] To tackle this problem an external validation was performed in the automated measurement tool.**(chapter 8)** Third, the responsibility in case of false outcomes that may or may not comprise patient safety is unclear. [22] To address these ethical considerations human control is necessary. In the case of the automated tumor volume measurement tool, a human validation of the volume is required. The AI-PREM is designed to support

human decision-making between quality improvement options points.(**chapters 6 and 7)** It is essential that clinicians place the outcomes into perspective, decide on which outcomes to act, and what action is appropriate. Incorporating the human factor in this way can mitigate the ethical risks when introducing data-driven solutions in health care delivery. Prospective controlled studies of data-driven technologies should evaluate the clinical implications of these ethical aspects.

Practical considerations include embedding data-driven tools in the clinicians' workflow in clinical practice and the applicability and clinical relevance of the outcomes provided by data-driven technologies. First, the starting point of new data-driven technologies should be a clinical problem or challenge. Clinicians are less likely to adopt developments that start with a technical solution and lack a clear perspective on the clinical issue that is addressed. Therefore, it is crucial to have multidisciplinary development teams including clinicians as end users, and technicians who can translate the clinical problem into a software tool that can provide helpful outcomes, which are presented in a user interface that is useful for clinicians and is understandable. [25] Second, the implementation of the tool in the clinical workflow should be smooth and easily accessible. [26] The existing administrative burden and highly complex and interrupting workflow of clinicians limited the use of completely new software tools. For smooth adoption new tools should be integrated with the existing software tools, such as electronic records or picture archiving communication systems (PACS). Development of the user interface is preferably performed in collaboration with the end-users. In addition, clinicians should be trained to use and assess data-driven technologies. [25]

10.4 Limitations of value-based healthcare

VBHC has spread in the medical world during the last decade. Although many elements of VBHC can be seen in previous quality of care initiatives, the comprehensive VBHC framework helps healthcare organizations transform to become more patient-centered and improve their quality by optimizing patient value.

Many healthcare quality improvement methods have emerged over the last two decades, and most of them disappeared after cycles of approximately five years. [27] Numerous methods, such as 'total quality management', 'continuous quality improvement', 'six sigma', and 'lean' have very similar fundamentals with different presentations and accents. Walshe coined this iterative development of quality improvement methods as "pseudo-innovation" or reinvention, which can be caused by both the financial incentives of quality improvement developers and implementation consultants and the willingness of users, clinicians, and hospital managers to obtain quick-fix quality improvement measures. [27] Pilot projects often result in promising outcomes which are not easily

reproduced in different settings. The quality improvement methods lack the proper empirical and experiential evidence, which is for example required when introducing new clinical innovations in practice. VBHC also has several characteristics of being the latest member of the pseudo-innovation quality improvement family. [28] When citations of the landmark papers about VBHC were analyzed, the rapidly increasing number of citation as well as the lack of understanding of the VBHC concept in the citing papers are suggestive of pseudo-innovation. [28]

In addition, the original concept of VBHC is described in general terms, which leave room for different interpretations. As explained in **chapter 1**, the way VBHC is adopted in different countries varies, as different components are highlighted, omitted, or newly introduced into the model. All these different interpretations limit the comparability and transferability of VBHC implementation at specific hospitals in specific countries. [29] However, this level of adaptability of VBHC can help to fit the VBHC concepts to the particular organizational and cultural context and improve the chance of successful implementation. [30]

One of the key differences in the interpretations of the VBHC concepts can be explained by the various ways in which 'value' has been interpreted. Porter et al. used value in a predominantly economic sense as outcomes divided by the costs, whereas in several European countries, a more moral sense of value has been adopted in which value is the good or desirable thing to do. [28, 31] This difference in perspective on value, together with the different financing structure of healthcare, probably explains the fact that the costs play a less prominent role in these European versions of VBHC. To achieve true value-based health care delivery, the outcomes measured should be divided by its costs. This enables the identification of inefficiencies and allows further cost-efficient improvement. [32] According to Porter and Kaplan, time-driven activity-based costing is the method of choice for determining the cost of healthcare. [33]

Besides the financial aspect, costs also comprise the burden of patients and the time invested by professionals to achieve the outcomes. Integrating VBHC requires a change in the organization of care. [34] Measuring the same core outcomes in every patient necessitates a change in the way of working of the healthcare providers, as they need to alter their data input in the electronic patient records. Not only are additional data required such as PROM and PREM results, but key parameters have to be filed as discrete data in order to retrieve them with ease from the EPD. Furthermore, integrating care across medical specialties also involves monthly meetings to discuss the organization of care and the aim, progress, and results of quality improvement initiatives. The effort

to achieve value-based vestibular schwannoma care, in other words, could very well increase the workload for healthcare professionals.

In addition, the measurement of patient-reported outcomes such as PROMs and PREMs requires an additional time investment from patients as well. The work of being a patient and the treatment burden of patients with chronic diseases can impact therapy adherence and quality of life, especially when the impact of the work on disease outcome is unclear.[35] The use of PROMs as part of VBHC provides valuable insight at the group level as described in this chapter. But the results should also be discussed on an individual level, with the patient in the outpatient clinic, otherwise, the risk of lower response rates exists. In addition, the increased burden for patients runs counter to some of the concepts of patient centeredness. This seems paradoxical since it is argued that VBHC promotes patient centeredness.[36] Therefore, the burden on patients must be weighed against the potential benefits in terms of patient centeredness.

The original VBHC concept as coined by Porter has a strong economic perspective with focus on competition, and centralization. Centralization is probably wise for rare conditions and complex treatments, such as vestibular schwannoma. However, there is evidence that this is not true for all conditions. A striking example is emergency care. In Denmark, this care has been concentrated and the number of hospitals with an emergency room has been halved. However, recent studies show that mortality did not decrease and length of stay and admissions remained the same.[37,38] Again, it can be argued that patient centeredness actually decreases due to, for example, longer travel times while quality does not clearly improve.

10.5 Future research

Several promising topics for future research emerge from this thesis. First, automated volume measurement paves the way for more accurate volume prediction. Using large quantities of clinical and radiological data may improve the currently poorly performing predictions of tumor growth. When more accurate prediction is possible, timely active treatment in progressive vestibular schwannomas may reduce the uncertainty patients with vestibular schwannoma are experiencing, as described in **chapter 5**, and change the decision-making process. Second, the intersection of patient-centeredness and value-based healthcare presents an intriguing research focus, exploring ways to optimize outcome measurements while mitigating the burden on patients. Third, investigating the cost elements of vestibular schwannoma care holds the potential to optimize value creation by identifying areas where efficiency gains can be made. Last, the effectiveness of multidisciplinary collaboration in enhancing overall value in the context of vestibular schwannoma care is a compelling area for further study, with the aim of clarifying the

specific mechanisms that contribute to increased value through collaborative efforts. These future research topics may increase our understanding and contribute to the continuous improvement of value based and data driven vestibular schwannoma care.

10.6 Conclusion and Implications

Part 1 of this thesis revealed that PROMs offer valuable insights at the group level, and the disease-specific questionnaire encompasses issues relevant to vestibular schwannoma patients. However, utilizing these outcomes in treatment decision-making and in the evaluation of vestibular schwannoma care presents challenges. The inherent nature of the disease—where the treatment is not aimed at symptom resolution but at averting potential severe complications—restricts the utility of PROMs for pre- and post-treatment evaluations. Furthermore, treatment decisions are primarily driven by tumor factors like size and progression. In contrast, symptom progression or a decline in quality of life only marginally influence these decisions. The fact that outcomes do not drive decisions does not imply a lack of patient-centeredness, as both qualitative research and patient experiences (**chapter 5 and 7**) affirm satisfaction with the current decision-making process.

Part 2 of this thesis focused on the value of data-driven technologies and how they can assist in creating value in vestibular schwannoma care. Close and intensive collaboration with data-scientists resulted in two tools that catalyze value creation by improving measurement accuracy (c**hapters 8 and 9**) and assisting continuous quality improvement based on patients' experiences (**chapter 6 and 7**) while reducing workload. Both tools have the potential to be used in other diseases.

In the realm of vestibular schwannoma care, the integration of value-based and data-driven methodologies brings forth useful elements that enhance overall quality, promote effective multidisciplinary collaboration, and leverage technology to improve the standard of care. While patient-reported outcomes may not directly dictate specific treatment decisions due to the inherent nature of the disease, their utility lies in providing contextual information at the group level. Long-term outcomes serve as insights for patients in their decision-making processes. Moreover, the incorporation of data-driven technologies not only enhances measurement precision but also facilitates continuous quality improvement and offers the potential to reduce workload.

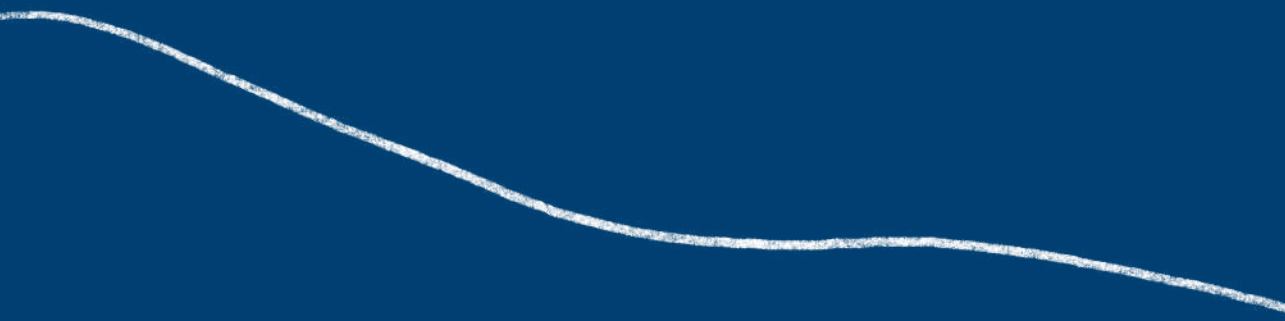
Recognizing that value-based and data-driven care is not universally applicable as a one size fit all package, it necessitates a tailored approach for each medical condition. The thoughtful selection of elements, specifically adapted to the unique characteristics of each disease, becomes paramount for optimizing patient value.

REFERENCES

1. Steinmann G, Van De Bovenkamp H, De Bont A, Delnoij D. Redefining value: a discourse analysis on value-based health care. *BMC Health Serv Res*. 2020;20(1)doi:10.1186/s12913-020-05614-7
2. Ramsden RT. The bloody angle: 100 years of acoustic neuroma surgery. *J R Soc Med*. 1995;88(8):464P-468P.
3. Torres Maldonado S, Naples JG, Fathy R, et al. Recent Trends in Vestibular Schwannoma Management: An 11-Year Analysis of the National Cancer Database. *Otolaryngology–Head and Neck Surgery*. 2019;161(1):137-143. doi:10.1177/0194599819835495
4. Chan SA, Marinelli JP, Hahs-Vaughn DL, Nye C, Link MJ, Carlson ML. Evolution in Management Trends of Sporadic Vestibular Schwannoma in the United States Over the Last Half-century. *Otol Neurotol*. 2021;42(2):300-305. doi:10.1097/mao.0000000000002891
5. Gauden A, Weir P, Hawthorne G, Kaye A. Systematic review of quality of life in the management of vestibular schwannoma. *J Clin Neurosci*. 2011;18(12):1573-1584. doi:10.1016/j.jocn.2011.05.009
6. Shaffer BT, Cohen MS, Bigelow DC, Ruckenstein MJ. Validation of a disease-specific quality-of-life instrument for acoustic neuroma. *The Laryngoscope*. 2010;120(8):1646-1654. doi:10.1002/lary.20988
7. Pruijn IMJ, van Heemskerken P, Kunst HPM, Tummers M, Kievit W. Patient-preferred outcomes in patients with vestibular schwannoma: a qualitative content analysis of symptoms, side effects and their impact on health-related quality of life. *Qual Life Res*. Oct 2023;32(10):2887-2897. doi:10.1007/s11136-023-03433-x
8. Jahoda M. *Employment and unemployment: A social-psychological analysis*. vol 1. Cambridge University Press; 1982.
9. Bartley M. Employment status, employment conditions, and limiting illness: prospective evidence from the British household panel survey 1991-2001. *J Epidemiol Community Health*. 2004;58(6):501-506. doi:10.1136/jech.2003.009878
10. Montgomery SM, Cook DG, Bartley MJ, Wadsworth ME. Unemployment pre-dates symptoms of depression and anxiety resulting in medical consultation in young men. *Int J Epidemiol*. 1999;28(1):95-100. doi:10.1093/ije/28.1.95 %J International Journal of Epidemiology
11. McGlothlin AE, Lewis RJ. Minimal Clinically Important Difference. *JAMA*. 2014;312(13):1342. doi:10.1001/jama.2014.13128
12. Sedaghat AR. Understanding the Minimal Clinically Important Difference (MCID) of Patient-Reported Outcome Measures. *Otolaryngology–Head and Neck Surgery*. 2019;161(4):551-560. doi:10.1177/0194599819852604
13. Black N, Varaganum M, Hutchings A. Relationship between patient reported experience (PREMs) and patient reported outcomes (PROMs) in elective surgery. *BMJ Quality & Safety*. 2014;23(7):534-542. doi:10.1136/bmjqs-2013-002707
14. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open*. 2013;3(1):e001570. doi:10.1136/bmjopen-2012-001570
15. Greaves F, Jha AK. Quality and the curate's egg. *BMJ Quality & Safety*. 2014;23(7):525-527. doi:10.1136/bmjqs-2014-002993
16. Coulter A, Locock L, Ziebland S, Calabrese J. Collecting data on patient experience is not enough: they must be used to improve care. *BMJ*. 2014;348(mar26 1):g2225-g2225. doi:10.1136/bmj.g2225

17. Andersen SB, Birkelund R, Andersen MØ, Carreon LY, Coulter A, Steffensen KD. Factors Affecting Patient Decision-making on Surgery for Lumbar Disc Herniation. *Spine (Phila Pa 1976)*. 2019;44(2):143-149. doi:10.1097/BRS.0000000000002763
18. Kunneman M, Montori VM, Shah ND. Measurement with a wink. *BMJ Quality & Safety*. 2017;26(10):849-851. doi:10.1136/bmjqs-2017-006814
19. Smulders YE. Meet je kwaliteit met tevredenheid? *Ned Tijdschr Geneeskd*. 2022;166:B1989
20. Ahmed F, Burt J, Roland M. Measuring Patient Experience: Concepts and Methods. *The Patient - Patient-Centered Outcomes Research*. 2014;7(3):235-241. doi:10.1007/s40271-014-0060-5
21. van de Langenberg R, de Bondt BJ, Nelemans PJ, Baumert BG, Stokroos RJ. Follow-up assessment of vestibular schwannomas: volume quantification versus two-dimensional measurements. *Neuroradiology*. Aug 2009;51(8):517-24. doi:10.1007/s00234-009-0529-4
22. Morley J, Floridi L. An ethically mindful approach to AI for health care. *The Lancet*. 2020/01/25/ 2020;395(10220):254-255. doi:https://doi.org/10.1016/S0140-6736(19)32975-7
23. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. *J Med Ethics*. 2021;47(12):e3-e3. doi:10.1136/medethics-2019-105860
24. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. Sep 2019;25(9):1337-1340. doi:10.1038/s41591-019-0548-6
25. Matheny ME, Whicher D, Thadaney Israni S. Artificial Intelligence in Health Care. *JAMA*. 2020;323(6):509. doi:10.1001/jama.2019.21579
26. Zech JR, Santomartino SM, Yi PH. Artificial Intelligence (AI) for Fracture Diagnosis: An Overview of Current Products and Considerations for Clinical Adoption, From the AJR Special Series on AI Applications. *American Journal of Roentgenology*. 2022:1-10. doi:10.2214/AJR.22.27873
27. Walshe K. Pseudoinnovation: the development and spread of healthcare quality improvement methodologies. *Int J Qual Health Care*. 2009;21(3):153-159. doi:10.1093/intqhc/mzp012
28. Fredriksson JJ, Ebbevi D, Savage C. Pseudo-understanding: an analysis of the dilution of value in healthcare. *BMJ Quality & Safety*. 2015;24(7):451. doi:10.1136/bmjqs-2014-003803
29. Van Staalduinen DJ, Van Den Bekerom P, Groeneveld S, Kidanemariam M, Stiggelbout AM, Van Den Akker-Van Marle ME. The implementation of value-based healthcare: a scoping review. *BMC Health Serv Res*. 2022;22(1)doi:10.1186/s12913-022-07489-2
30. Colldén C, Hellström A. Value-based healthcare translated: a complementary view of implementation. *BMC Health Serv Res*. 2018;18(1)doi:10.1186/s12913-018-3488-9
31. Hazelzet JA, Thor J, Andersson Gäre B, et al. Value-based healthcare's blind spots: call for a dialogue. *F1000Research*. 2021;10:1314. doi:10.12688/f1000research.75578.1
32. Keel G, Savage C, Rafiq M, Mazzocato P. Time-driven activity-based costing in health care: A systematic review of the literature. *Health Policy*. Jul 2017;121(7):755-763. doi:10.1016/j.healthpol.2017.04.013
33. Kaplan DM. Using Time-Driven ActivityBased Costing to Identify Value Improvement Opportunities in Healthcare. *Healthcare management*. 2014;59(6)
34. Van Harten W. Turning teams and pathways into integrated practice units: Appearance characteristics and added value. *International Journal of Care Coordination*. 2018;21(4):113-116. doi:10.1177/2053434518816529
35. Tran V-T, Barnes C, Montori VM, Falissard B, Ravaud P. Taxonomy of the burden of treatment: a multi-country web-based qualitative study of patients with chronic conditions. *BMC Med*. 2015;13(1)doi:10.1186/s12916-015-0356-x

36. Kidanemariam M, Pieterse AH, van Staalduinen DJ, Bos WJW, Stiggelbout AM. Does value-based healthcare support patient-centred care? A scoping review of the evidence. *BMJ Open*. Jul 10 2023;13(7):e070193. doi:10.1136/bmjopen-2022-070193
37. Flojstrup M, Bogh SBB, Bech M, Henriksen DP, Johnsen SP, Brabrand M. Mortality before and after reconfiguration of the Danish hospital-based emergency healthcare system: a nationwide interrupted time series analysis. *BMJ Qual Saf*. Apr 2023;32(4):202-213. doi:10.1136/bmjqs-2021-013881
38. Bogh SB, Fløjstrup M, Möller S, et al. Intended and unintended changes in length of stay following reconfiguration of emergency care departments. *Int J Qual Health Care*. Feb 5 2021;33(1) doi:10.1093/intqhc/mzab008



CHAPTER 11

Summary

Vestibular schwannoma care encompasses several challenging aspects regarding decision making, organization and evaluation of care. This rare and benign intracranial tumor requires multidisciplinary care to provide diverse treatment options. The disease, treatment, and its sequelae impact patients' daily life, and the timing of active treatment is delicate. This complexity makes vestibular schwannoma care suited to organize care delivery to promote continuous improvement of quality.

In this thesis, principles of value-based and data-driven healthcare were studied in the context of vestibular schwannoma care. The outcomes of vestibular schwannoma care are relevant to understand the patients' perspective on the disease. Furthermore, two artificial intelligence based tools that can assist quality improvement and evaluation were developed and assessed.

Background

Chapter 1 provides background information on vestibular schwannoma, value-based healthcare, and data driven care. Vestibular schwannomas are benign tumors arising from Schwann cells of the vestibulocochlear nerve. Over the years the incidence has increased, most likely due to improved diagnostic methods. Treatment options include active surveillance, surgery, and radiotherapy, each with its own risks and outcomes. The aim of treatment is to prevent future serious complications due to compression of a growing tumor on vital structures. The presenting symptoms cannot be alleviated by treatment and are likely to worsen in all three treatment modalities. Decision-making involves weighing tumor and patient factors, including current symptoms, tumor size and progression, necessitating shared decision making between patients and clinicians.

Furthermore, the development of value-based healthcare is described. In the last decades, value-based healthcare has emerged rapidly as a solution to improve patient value, which is defined as health outcomes relevant for patients divided by the cost and burden needed to achieve these outcomes. Value-based healthcare has been developed in the United States of America and focuses on the competition between healthcare providers. In Europe and specifically in the Netherlands value-based healthcare is aimed at empowering patients and improving patient-centered care delivery. Value-based healthcare in the Netherlands consists of four components. First, care should be organized in care teams treating a specific disease. Second, an outcome set with clinical and patient-reported outcomes should be defined and measured in every patient. Third, shared decision-making is essential to deliver truly patient-centered care. Fourth, data technology should assist the individual and group level analysis of all collected data and should be used to continuously improve the quality of care.

Data-driven technologies are essential for both analyzing all collected patient data and improving quality of care while reducing clinician's workload. Challenges include integrating data sources, and collaboration between clinicians and data scientists. Artificial intelligence applications aim to assist in diagnosis, predict outcomes, and analyze large datasets. Despite these promising objectives clinical adoption remains limited due to integration complexities, data quality issues, and liability discussions.

Value-based vestibular schwannoma care

Long-term quality of life of vestibular schwannoma patients were evaluated in **chapter 2.** A cohort of vestibular schwannoma patients which participated in quality of life research in 2014 was approached for participation. In total 536 patients completed the questionnaires and together with the 2014 results a longitudinal analysis was performed. On average the long-term quality of life of patients was comparable between active surveillance, surgery, and radiotherapy. The quality of life was stable over time. Patients requiring salvage therapy after initial therapy failure showed lower quality of life scores. In addition, the study described in **chapter 3** showed that there were no differences in employment rates of vestibular schwannoma patients and an age matched Dutch general population group. Neither did treatment strategies impact working hours, employment rates or absenteeism.

The response rate of patient reported outcomes was analyzed in **chapter 4.** In this study, patients received the questionnaires by post or email. Regular mail delivery had the best response rates, however, it is more time consuming in distribution and digitalization. Email delivery had the lowest response rates and a hybrid delivery method in which patients receive a letter by regular mail with a code to access the survey electronically scored in between email and post delivery. Therefore, hybrid delivery might be the best of both worlds, with a relatively high response rate without the workload of digitalization.

Chapter 5 described a qualitative study on factors that influence patient decision making. Eighteen patients were interviewed about their treatment decisions. Besides well-known medical factors such as tumor characteristics and physicians' recommendations, patient related factors also impacted the decision making. Anxiety and experiences of relatives with certain treatment modalities (surgery or radiotherapy) influenced the decision making, as did non-medical factors such as time to return to work, or ways of dealing with the uncertainty of treatment outcomes. Addressing these factors during consultation can improve shared decision making.

Data-driven vestibular schwannoma care

Patient experiences are important indicators of quality of care. They can be measured using patient-reported experience measures (PREMs), usually in the form of questionnaires. However close ended PREMS tend to show a ceiling effect and if they show negative experiences, these are often too generic to translate them into action points for quality improvement. Open ended questions provide more context dependent answers that can be translated to points for quality improvement more easily. However, the analysis of open text comments is very time consuming. A newly developed artificial intelligence based PREM using open questions combined with an automated analysis was reported in **chapter 6 and 7.**

Chapter 6 describes the development of an automated analysis of open-ended PREM questions. A new questionnaire (AI-PREM) was developed to facilitate automated analysis. The sentiment of the open-ended answers was classified as positive, negative, or neutral. In addition, the tool clustered the answers which contained information on the same subjects. The final interpretation is still performed by the clinicians. **Chapter 7** showed that AI-PREM results lead to more relevant action points for quality improvement compared to results of a conventional close ended PREM. Even patients who in general had an excellent patient experience provided valuable suggestions for quality improvement in the AI-PREM.

Chapters 8 and 9 showed the development of an automated measurement tool of vestibular schwannoma on MRI. In clinical practice the largest extrameatal diameter is often used to measure the tumor size. Although these two-dimensional measurements are easy to obtain, there is a considerable intra- and interobserver variability. Based on the measurement error a 2 mm cut-off point is used to determine tumor growth on two consecutive scans. Volume measurements are known to be more accurate but are more time demanding since the tumor should be delineated on every slice of the MRI-scan.

In **chapter 8,** we have trained an algorithm to automatically measure tumor volumes on T1 post contrast and high resolution T2 sequences. The algorithm was trained on scans acquired in 37 different hospitals and 12 different MRI scanners. The algorithm could accurately delineate tumors and make a distinction between intra- and extrameatal tumor parts. The tool performed comparably to human delineation in 87-93% of the cases. External validation in a publicly available data set showed consistent results. In **chapter 9** the measurement of two-dimensional extrameatal diameters was evaluated. The automated tool could measure these diameters as well as human measurements. Furthermore, the tool was able to accurately detect tumor growth on consecutive scans.

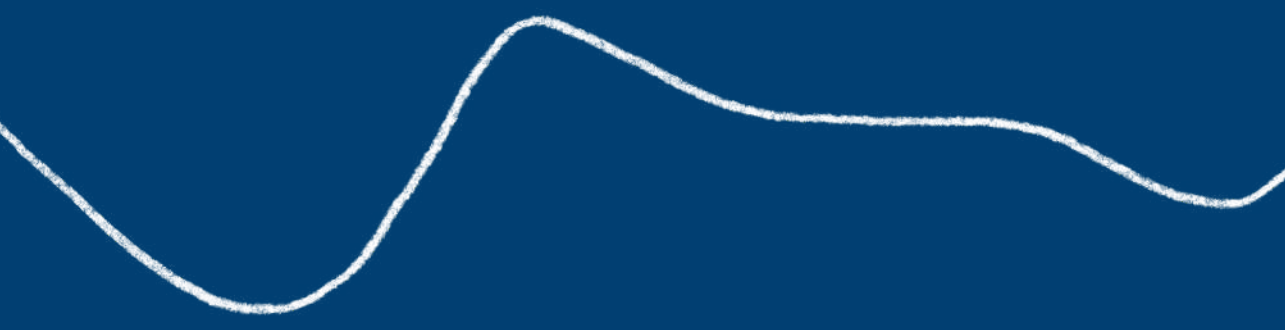
Discussion and conclusions

Long term results show no differences in quality of life and employment rates between patients who underwent the different treatment modalities. These results on group level can assist physicians to counsel patients during their decision-making process. The use of quality of life PROMs for evaluation of care, which is advocated in the value-based healthcare paradigm, has limitations in vestibular schwannoma care. The inherent nature of the disease—where the treatment is not aimed at symptom resolution but at averting potential severe complications—restricts the utility of PROMS for pre- and post-treatment evaluations. Furthermore, treatment decisions are primarily driven by tumor factors like size and progression.

The use of the two developed data driven technologies can facilitate quality improvement in vestibular schwannoma care. Patients' experiences can assist continuous quality improvement while reducing the workload due to the automated analysis. In addition, more accurate measurement of tumors can help decisions making, while reducing time needed for performing those measurements. In both cases, human validation of interpretation of the automated results is still essential for clinical practice. Furthermore, incorporating the data driven tools seamlessly in the current workflow is important to enhance clinical adoption.

Future research can be aimed at using automated volume measurement to study more precise growth prediction. These predictions may eventually lead to timelier active treatment in progressive tumors.

The integration of value-based and data-driven methodologies brings forth useful elements that enhance overall quality, promote effective multidisciplinary collaboration, and leverage technology to improve vestibular schwannoma care. However, value-based and data-driven care is not universally applicable as a one size fits all package, it necessitates a tailored approach for each medical condition to achieve true value optimization.



CHAPTER 12

Nederlandse samenvatting

Zorg rond vestibularis schwannomen omvat verschillende uitdagende aspecten met betrekking tot besluitvorming, organisatie en evaluatie van de zorg. De zeldzame en goedaardige intracraniële tumor vereist multidisciplinaire zorg die verschillende behandelingsopties biedt. De ziekte, de behandeling en de gevolgen hebben invloed op het dagelijks leven van patiënten en de timing van behandeling luistert nauw. Deze complexiteit maakt de aandoening uitermate geschikt voor de organisatie van zorg in een multidisciplinair zorgpad gericht op voortdurende kwaliteitsverbetering.

In dit proefschrift werden principes van waarde- en datagedreven gezondheidszorg bestudeerd in de context van vestibularis schwannoom zorg. De uitkomsten van vestibularis schwannoom zorg zijn relevant om het perspectief van de patiënt op de ziekte te begrijpen. Verder werden twee op kunstmatige intelligentie gebaseerde hulpmiddelen ontwikkeld en getest die gericht zijn op kwaliteitsverbetering en continue evaluatie van zorg.

Achtergrond

Hoofdstuk 1 geeft achtergrondinformatie over vestibularis schwannomen, waardegedreven zorg en datagedreven zorg. Vestibularis schwannomen zijn goedaardige tumoren die ontstaan uit Schwann cellen van de nervus vestibulocochlearis. In de loop der jaren is de incidentie toegenomen, waarschijnlijk door verbeterde diagnostiek. Behandelopties zijn observatie met MRI-scans, chirurgie en radiotherapie, elk met hun eigen risico's en resultaten. Het doel van de behandeling is om toekomstige ernstige complicaties door compressie van een groeiende tumor op vitale structuren te voorkomen. Tumorgroei is daarom een essentiële factor. Veel tumoren groeien niet of heel traag. Na de diagnose blijft 60% van de tumoren stabiel, terwijl 40% groeit, soms pas na meerdere jaren.

Behandeling is niet gericht op het verlichten van symptomen. Sterker nog, na behandeling zullen symptomen vaak verergeren. Ook bij afwachtend beleid kunnen symptomen in de loop der tijd toenemen. Bij de besluitvorming moeten factoren met betrekking tot de tumor en de patiënt worden afgewogen, waaronder de huidige symptomen en de grootte en progressie van de tumor.

Verder wordt de ontwikkeling van waardegedreven zorg beschreven. In de afgelopen decennia is waardegedreven zorg snel naar voren gekomen als methode om continu de kwaliteit van zorg te verbeteren. Het doel is om waarde voor waarde voor de patiënte te optimaliseren. Deze waarde wordt gedefinieerd als gezondheidsresultaten die relevant zijn voor patiënten gedeeld door de kosten en lasten die nodig zijn om deze resultaten te bereiken. Waardegedreven zorg is ontwikkeld in de Verenigde Staten en richtte zich oorspronkelijk op de concurrentie tussen zorgverleners. In Europa en specifiek in Nederland is waardegedreven zorg gericht op het mondiger maken van patiënten en het ver-

beteren van patiëntgerichte zorgverlening. Waardegedreven zorg in Nederland bestaat uit vier componenten. Ten eerste moet de zorg worden georganiseerd in zorgteams die een specifieke ziekte behandelen. Ten tweede moet bij elke patiënt een selectie van uitkomsten met klinische en patiëntgerapporteerde resultaten worden gedefinieerd en gemeten. Ten derde is gedeelde besluitvorming essentieel om daadwerkelijk patiëntgerichte zorg te leveren. Ten vierde moet datatechnologie helpen bij de analyse van alle verzamelde gegevens op individueel- en groepsniveau en gebruikt worden om de kwaliteit van de zorg continu te verbeteren.

Data gestuurde technologieën zijn essentieel voor zowel het analyseren van alle verzamelde patiëntgegevens als het verbeteren van de kwaliteit van de zorg, terwijl de werkdruk van artsen afneemt. Uitdagingen zijn onder andere de integratie van gegevensbronnen en samenwerking tussen clinici en gegevenswetenschappers. Kunstmatige intelligentietoepassingen hebben als doel te helpen bij het stellen van diagnoses, het voorspellen van uitkomsten en het analyseren van grote datasets. Ondanks deze veelbelovende doelen blijft de klinische toepassing tot op heden beperkt vanwege de complexe integratie, problemen met gegevenskwaliteit en discussies over aansprakelijkheid.

Waardegedreven vestibularis schwannoom zorg

In hoofdstuk 2 wordt de kwaliteit van leven vestibularis schwannoom patiënten op lange termijn geëvalueerd. Een cohort van vestibularis schwannoom patiënten dat in 2014 deelnam aan onderzoek naar kwaliteit van leven werd opnieuw benaderd voor deelname. In totaal vulden 536 patiënten de vragenlijsten in en samen met de resultaten van 2014 werd een longitudinale analyse uitgevoerd. Gemiddeld was de kwaliteit van leven van patiënten op de lange termijn vergelijkbaar tussen observatie, chirurgie en radiotherapie. De kwaliteit van leven was stabiel in de loop van de tijd. Patiënten die laatste lijns therapie nodig hadden na falen van de initiële therapie, scoorden lager op kwaliteit van leven. Daarnaast toont het onderzoek beschreven in **hoofdstuk 3** aan dat er geen verschillen waren in arbeidsparticipatie van vestibularis schwannoom patiënten en leeftijdsgenoten in de algemene Nederlandse bevolkingsgroep. Behandelingsstrategieën hadden ook geen invloed op werktijden, arbeidsparticipatie of ziekteverzuim.

De respons van door patiënten gerapporteerde uitkomsten wordt geanalyseerd in **hoofdstuk 4**. In dit onderzoek ontvingen patiënten de vragenlijsten per post of e-mail. Postbezorging had de beste respons, maar is tijdrovender in distributie en digitalisering. Bezorging per e-mail had de laagste respons en een hybride bezorgmethode waarbij patiënten een brief per post ontvingen met een code om elektronisch toegang te krijgen tot de enquête scoorde tussen bezorging per e-mail en per post in. Daarom zou hybride

bezorging het beste van twee werelden kunnen zijn, met een relatief hoge respons zonder de werklast van digitalisering.

Hoofdstuk 5 beschrijft een kwalitatief onderzoek naar factoren die de besluitvorming van patiënten beïnvloeden. Achttien patiënten werden geïnterviewd over hun behandelbeslissingen. Naast medische factoren zoals tumorkenmerken en aanbevelingen van artsen, waren ook patiëntgerelateerde factoren van invloed op de besluitvorming. Angst en ervaringen van familieleden met bepaalde behandelmethoden (chirurgie of radiotherapie) beïnvloedden de besluitvorming, net als niet-medische factoren zoals de tijd tot terugkeer op het werk of manieren van omgaan met de onzekerheid over de uitkomst van de behandeling. Begrip voor deze factoren tijdens het consult kan de gedeelde besluitvorming verbeteren.

Datagedreven vestibularis schwannoom zorg

Patiëntervaringen zijn belangrijke indicatoren voor de kwaliteit van zorg. Ze kunnen worden gemeten met behulp van door patiënten gerapporteerde ervaringsmetingen (PREMs), meestal in de vorm van vragenlijsten. PREMs met gesloten vragen hebben echter de neiging een plafondeffect te vertonen en als ze negatieve ervaringen laten zien, zijn deze vaak te algemeen om ze te vertalen in actiepunten voor kwaliteitsverbetering. Open vragen geven meer contextafhankelijke antwoorden die gemakkelijker vertaald kunnen worden naar punten voor kwaliteitsverbetering. De analyse van deze vrije tekst antwoorden is echter zeer tijdrovend. In **hoofdstuk 6 en 7** wordt verslag gedaan van een nieuw ontwikkelde, op kunstmatige intelligentie gebaseerde PREM die gebruik maakt van open vragen in combinatie met een geautomatiseerde analyse.

Hoofdstuk 6 beschrijft de ontwikkeling van een geautomatiseerde analyse van open PREM-vragen. Er werd een nieuwe vragenlijst (AI-PREM) ontwikkeld om automatische analyse mogelijk te maken. Het sentiment van de open antwoorden werd geclassificeerd als positief, negatief of neutraal. Daarnaast clusterde de applicatie de antwoorden die informatie over dezelfde onderwerpen bevatten. De uiteindelijke interpretatie werd nog steeds uitgevoerd door de clinici. **Hoofdstuk 7** toont aan dat AI-PREM resultaten leiden tot relevantere actiepunten voor kwaliteitsverbetering in vergelijking met de resultaten van een conventionele PREM. Zelfs patiënten die over het algemeen een uitstekende patiëntervaring hadden, gaven waardevolle suggesties voor kwaliteitsverbetering in de AI-PREM.

Hoofdstuk 8 en 9 tonen de ontwikkeling van een geautomatiseerd meetinstrument voor een vestibularis schwannoom op een MRI-scan. In de klinische praktijk wordt vaak de grootste extrameatale diameter gebruikt om de tumorgrootte te meten. Hoewel deze

tweedimensionale metingen eenvoudig te verkrijgen zijn, is er een aanzienlijke intra- en interobserver variabiliteit. Op basis van de meetfout wordt een afkappunt van 2 mm gebruikt om de tumorgroei op twee opeenvolgende scans te bepalen. Van volumemetingen is bekend dat ze nauwkeuriger zijn, maar ze vergen meer tijd omdat de tumor op elke coupe van de MRI-scan moet worden ingetekend.

Hoofdstuk 8 beschrijft de ontwikkeling van een algoritme om automatisch tumorvolumes te meten op T1 post contrast en hoge resolutie T2 sequenties. Het algoritme werd getraind op scans die waren verkregen in 37 verschillende ziekenhuizen en afkomstig waren van 12 verschillende scanners. Het algoritme kon nauwkeurig tumoren intekenen en onderscheid maken tussen intra- en extrameatale tumor onderdelen. De applicatie presteerde in 87-93% van de gevallen vergelijkbaar met menselijke metingen. Externe validatie in een openbaar beschikbare dataset toonde consistente resultaten. In **hoofdstuk 9** wordt de automatische meting van tweedimensionale extrameatale diameters geëvalueerd. Het geautomatiseerde hulpmiddel kon deze diameters net zo goed meten als mensen. Bovendien was het hulpmiddel in staat om tumorgroei nauwkeurig te detecteren op opeenvolgende scans.

Discussie en conclusie

De resultaten op de lange termijn laten geen verschillen zien in kwaliteit van leven en arbeidsparticipatie tussen patiënten die verschillende behandelmethoden ondergingen. Deze resultaten op groepsniveau kunnen artsen helpen om patiënten te begeleiden tijdens hun besluitvormingsproces. Het gebruik van door de patiënt gerapporteerde uitkomsten voor de evaluatie van zorg wordt bepleit in het waardegedreven zorg paradigma. In het kader van zorg rond vestibularis schwannomen heeft het gebruik van zulke kwaliteit van leven uitkomsten beperkingen. De inherente aard van de ziekte - waarbij de behandeling niet gericht is op het verlichten van symptomen maar op het voorkomen van mogelijke ernstige complicaties - beperkt de bruikbaarheid van de uitkomsten voor evaluaties van voor en na de behandeling. Bovendien worden beslissingen over behandeling voornamelijk bepaald door tumorfactoren zoals grootte en progressie.

Het gebruik van de twee ontwikkelde data gestuurde technologieën kan kwaliteitsverbetering in de zorg rond vestibularis schwannomen vergemakkelijken. Ervaringen van patiënten kunnen bijdragen aan continue kwaliteitsverbetering, terwijl de werkdruk door de geautomatiseerde analyse afneemt. Daarnaast kunnen nauwkeurigere metingen van tumoren helpen bij het nemen van beslissingen, terwijl de tijd die nodig is voor het uitvoeren van die metingen wordt verminderd. In beide gevallen is menselijke validatie van de interpretatie van de geautomatiseerde resultaten nog steeds essentieel voor de klinische praktijk. Bovendien is het belangrijk om de data gestuurde applica-

ties naadloos in de huidige werkzaamheden te integreren om de klinische adoptie te verbeteren.

Toekomstig onderzoek kan zich richten op het gebruik van geautomatiseerde volumemeting om nauwkeurigere groeivoorspellingen mogelijk te maken. Deze voorspellingen kunnen uiteindelijk leiden tot een tijdigere actieve behandeling bij progressieve tumoren en maat gemaakte follow-up met mogelijk minder scans.

De integratie van op waardegedreven en data gestuurde methodologieën brengt nuttige elementen voort die de algehele kwaliteit verbeteren, effectieve multidisciplinaire samenwerking bevorderen en technologie inzetten om de zorg voor vestibularis schwannomen te verbeteren. Waarde- en datagedreven zorg is echter niet universeel toepasbaar als een alles-in-één pakket, het vereist een aanpak op maat voor elke medische aandoening om waarde optimalisatie te bereiken.

CHAPTER 13

Appendices

List of publications
Curriculum vitae
Dankwoord

LIST OF PUBLICATIONS

Neve OM, Soulier G, Hendriksma M, van der Mey AGL, van Linge A, van Benthem PPG, Hensen EF, Stiggelbout AM
Patient-reported factors that influence the vestibular schwannoma treatment decision: a qualitative study.
Eur Arch Otorhinolaryngol. 2020;doi:10.1007/s00405-020-06401-0

Neve OM, Boerman JA, van den Hout WB, Briaire JJ, van Benthem PPG, Frijns JHM.
Cost-benefit Analysis of Cochlear Implants: A Societal Perspective.
Ear Hear. Mar 4 2021;doi:10.1097/aud.0000000000001021

Neve OM, Jansen JC, Van Der Mey AGL, Koot RW, De Ridder M, Van Benthem PPG, Stiggelbout AM, Hensen EF
The impact of vestibular schwannoma and its management on employment.
Eur Arch Otorhinolaryngol. 2021;doi:10.1007/s00405-021-06977-1

Neve OM, van Benthem PPG, Stiggelbout AM, Hensen EF
Response rate of patient reported outcomes: the delivery method matters.
BMC Med Res Methodol. Oct 22 2021;21(1):220. doi:10.1186/s12874-021-01419-2

Neve OM, Jansen JC, Koot RW, de Ridder M, van Benthem PPG, Stiggelbout AM, Hensen EF
Long-Term Quality of Life of Vestibular Schwannoma Patients: A Longitudinal Analysis.
Otolaryngol Head Neck Surg. 2022:019459982210885. doi:10.1177/01945998221088565

Van Buchem MM, **Neve OM**, Kant IMJ, Steyerberg EW, Boosman H, Hensen EF.
Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (AI-PREM).
BMC Med Inform Decis Mak. 2022;22(1)doi:10.1186/s12911-022-01923-5

Neve OM*, Chen Y*, Tao Q, Romeijn SR, de Boer NP, Grootjans W, Kruit MC, Lelieveldt BPF, Jansen JC, Hensen EF, Verbist BM, Staring M
Fully Automated 3D Vestibular Schwannoma Segmentation with and without Gadolinium-based Contrast Material: A Multicenter, Multivendor Study.
Radiology: Artificial Intelligence. 2022;4(4):e210300. doi:10.1148/ryai.210300

Neve OM, Romeijn SR, Chen Y, Nagtegaal L, Grootjans W, Jansen JC, Staring M, Verbist BM, Hensen EF
Automated 2-Dimensional Measurement of Vestibular Schwannoma: Validity and Accuracy of an Artificial Intelligence Algorithm.
Otolaryngol Head Neck Surg. Dec 2023;169(6):1582-1589. doi:10.1002/ohn.470

Fuentealba-Bassaletti C, **Neve OM**, van Esch BF, Jansen JC, Koot RW, van Benthem PPG, Hensen EF
Vestibular Complaints Impact on the Long-Term Quality of Life of Vestibular Schwannoma Patients.
Otol Neurotol. 2023;44(2) doi: 10.1097/MAO.0000000000003773

Fuentealba-Bassaletti C, **Neve OM**, van Benthem PP, Hensen EF.
Reply to Letter to the Editor: "Vestibular Complaints Impact on the Long-Term Quality of Life of Vestibular Schwannoma Patients".
Otol Neurotol. Aug 1 2023;44(7):748. doi:10.1097/mao.0000000000003944

Neve OM, van Buchem MM, Kunneman M, van Benthem PPG, Boosman H, Hensen EF.
The added value of the artificial intelligence patient-reported experience measure (AI-PREM tool) in clinical practise: Deployment in a vestibular schwannoma care pathway.
PEC Innov. Dec 15 2023;3:100204. doi:10.1016/j.pecinn.2023.100204

Chen Y, Staring M, **Neve OM**, Romeijn SR, Hensen EF, Verbist BM, Wolterink JM, Tao Q
CoNeS: Conditional neural fields with shift modulation for multi-sequence MRI translation. *Machine Learning for Biomedical Imaging*. 2024;2(Special Issue for Generative Models):657-685. doi:https://doi.org/10.59275/j.melba.2024-d61g

Kidanemariam M, **Neve OM**, van den Heuvel I, Douz S, Hensen EF, Stiggelbout A.M., Pieterse A.H.
Patient-reported outcome measures in value-based healthcare: A multiple methods study to assess patient-centredness.
Patient Educ Couns. 2024 Mar 7:125:108243. doi: 10.1016/j.pec.2024.108243

CURRICULUM VITAE

Olaf Maarten Neve was born on 25 September 1993 in Hoorn. He completed his secondary education (Gymnasium) at the Openbare Scholengemeenschap West Friesland in Hoorn. Subsequently, he pursued his medical studies at Leiden University, earning his Bachelor's degree in 2014 and Master's degree in 2018. In the same year, he also achieved a Master's in Healthcare Management from the Erasmus School of Health Policy and Management at Erasmus University in Rotterdam.

He started working as a medical doctor at the surgery department of the Haaglanden Medisch Centrum in 2018. In 2019, he collaborated with Dr. Erik Hensen to develop a PhD proposal focusing on value-based vestibular schwannoma care within the Otorhinolaryngology Department. This proposal received approval and support from the strategic fund of Leiden University Medical Center (LUMC). He initiated his PhD journey in June 2019, working under the guidance of Professor van Benthem, Professor Stiggelbout, and Dr. Hensen.

Throughout his doctoral studies he joined the board of the national association representing residents in the Netherlands (De Jonge Specialist). In 2022 he started his residency in otorhinolaryngology at the LUMC.

DANKWOORD

Een promotietraject is een lange hobbelige weg, met hier en daar een doodlopend zijpad. Gelukkig hoef je niet hem niet in je eentje af te leggen. Tijdens mijn promotie heb ik de mogelijkheid gehad om met veel collega's van verschillende vakgebieden te kunnen samenwerken. De verschillende de achtergronden boden verrassende perspectieven en leiden tot vernieuwende ideeën. Zonder deze samenwerking was onmogelijk om dit traject tot een goed einde te kunnen brengen. Als ik iets geleerd heb, dan is het dat de complementaire expertises leiden tot betere resultaten en oplossingen die daadwerkelijk hun weg naar de kliniek vinden. Een aantal mensen wil ik in het bijzonder bedanken voor hun bijdrage aan dit proefschrift.

Allereerst mijn promotoren en copromotor. Professor van Benthem, Peter Paul, bedankt voor de coördinerende rol als promotor. Op enige afstand heb je het overzicht en de grote lijnen bewaakt en heb je met een kritische blik ook inhoudelijk de artikelen verbeterd.

Professor Stiggelbout, Anne, in onze maandelijkse besprekingen heb ik veel kunnen leren, zowel van jouw ervaring als klinisch epidemioloog als van jouw uitgebreide trackrecord met het begeleiden van promovendi. Altijd wist je oplossingen aan te dragen of suggesties te doen voor samenwerking.

Dr. Hensen, Erik, we zijn dit promotietraject begonnen met een voorstel op een A4'tje, waarna het daadwerkelijke traject zich gedurende de jaren daarna heeft vormgegeven door de voortdurende discussie die we hebben gevoerd. Dankzij jouw scherpe pen, wetenschappelijke ervaring en oog voor het klinische perspectief heb je een groot aandeel gehad in dit proefschrift en hebben we vaak als duo de resultaten kunnen presenteren binnen en buiten het LUMC.

Naast mijn de begeleiders van het promotie traject mag Andel van der Mey ook niet ontbreken. Jij hebt mijn interesse voor de KNO aangewakkerd en vervolgens ook nog wetenschappelijke aspiratie geïnitieerd, door mij als wetenschapsstudent bij het onderzoek van Géke Soulier onder te brengen. Jouw enthousiasme en kwaliteit om mensen samen te brengen is de motor achter het brughoektumoronderzoek in het LUMC.

Alle medische specialisten van het Schedelbasis centrum Leiden, Jeroen, Heiko, Radboud en Mischa bedankt voor jullie hulp bij patiënten includeren en leveren van input voor de artikelen. Waarschijnlijk nog veel belangrijker voor het onderzoek was de rol van Angela van Eijk, als case-manager een spin in het web die voor alle brughoektumor on-

derzoekers een steun en toeverlaat is en die onvermoeid informed-consentformulieren vergaard.

Speciale dank aan de collega's bij CAIRE-lab en directoraat kwlaiteit en patientveiligheid, Marieke van Buchem en Hileen Boosman, waarmee we samen de AI-PREM hebben ontwikkeld. Dankzij de mooie samenwerking hebben we van de grond af iets moois weten op te bouwen.

Ook veel dank voor de collega's, Marius, Yunjie, Stephan, Berit en Willem, die van uit de afdeling radiologen en laboratorium klinische experimentele beeldbewerking hebben samengewerkt aan het automatisch meten van brughoektumoren. Het was mooi om samen een van de eerste AI-projecten in het LUMC op te bouwen en nu zelfs onder leiding van Larissa naar de kliniek te brengen.

Alle collega onderzoekers bij KNO en medische besliskunde dank voor jullie sociale, mentale en intellectuele support. In het bijzonder de mede brughoektumor onderzoekers Nick, Kim, Constanza en Jules. Het was waardevol om met jullie te sparren en samen naar de brughoekcongressen te reizen. En vooral ook mijn kamergenoot op H5 Juliëtta, wie wil nou niet een klinische epidemioloog die al je statistiek vragen kan beantwoorden als kamergenoot?

Alle collega a(n)ios bij de KNO bedankt voor de collegiale sfeer en de bereidheid elkaar te helpen, dit maakt het werken elke dag plezierig. We zijn samen een goed gevarieerd team.

Nick en Lars, mijn paranifmen, ik waardeer jullie steun gedurende het traject. Fijn dat ik op jullie kan rekenen tijdens de verdediging en de activiteiten daaromheen.

William en Anneke, lieve ouders, en Karlijn bedankt voor jullie steun gedurende het traject, de blijvende interesse en nauwlettende controle op de voortgang van het onderzoek.

Jade, jij bent mijn belangrijkste steunpilaar en ik heb geluk dat jij aan mijn zijde staat.